An evaluation of the performance of chemistry transport models by comparison with research aircraft observations. Part 1: Concepts and overall model performance

D. Brunner, J. Staehelin, H. L. Rogers, M. O. Köhler, J. A. Pyle, D. Hauglustaine, Line Jourdain, T. K. Berntsen, M. Gauss, I.S.A. Isaksen, et al.

Atmos. Chem. Phys., 3, 1609–1631, 2003
www.atmos-chem-phys.org/acp/3/1609/

Atmospheric Chemistry and Physics

An evaluation of the performance of chemistry transport models by comparison with research aircraft observations. Part 1: Concepts and overall model performance

D. Brunner[1], J. Staehelin[1], H. L. Rogers[2], M. O. Köhler[2], J. A. Pyle[2], D. Hauglustaine[3], L. Jourdain[4], T. K. Berntsen[5], M. Gauss[5], I. S. A. Isaksen[5], E. Meijer[6], P. van Velthoven[6], G. Pitari[7], E. Mancini[7], V. Grewe[8], and R. Sausen[8]

[1]Institute for Atmospheric and Climate Science, ETH Zürich, Switzerland
[2]Centre for Atmospheric Science, Cambridge University, UK
[3]Laboratoire des Sciences du Climat et de L'Environnement, Gif-sur-Yvette, France
[4]Service d'Aéronomie, Paris, France
[5]Department of Geophysics, University of Oslo, Norway
[6]Section of Atmospheric Composition, Royal Netherlands Meteorological Institute, The Netherlands
[7]Dipartimento di Fisica, Università L'Aquila, Italy
[8]Institut für Physik der Atmosphäre, DLR, Germany

Received: 11 February 2003 – Published in Atmos. Chem. Phys. Discuss.: 15 May 2003
Revised: 18 August 2003 – Accepted: 31 August 2003 – Published: 6 October 2003

Abstract. A rigorous evaluation of five global Chemistry-Transport and two Chemistry-Climate Models operated by several different groups in Europe, was performed. Comparisons were made of the models with trace gas observations from a number of research aircraft measurement campaigns during the four-year period 1995–1998. Whenever possible the models were run over the same four-year period and at each simulation time step the instantaneous tracer fields were interpolated to all coinciding observation points. This approach allows for a very close comparison with observations and fully accounts for the specific meteorological conditions during the measurement flights. This is important considering the often limited availability and representativity of such trace gas measurements. A new extensive database including all major research and commercial aircraft measurements between 1995 and 1998, as well as ozone soundings, was established specifically to support this type of direct comparison. Quantitative methods were applied to judge model performance including the calculation of average concentration biases and the visualization of correlations and RMS errors in the form of so-called Taylor diagrams. We present the general concepts applied, the structure and content of the database, and an overall analysis of model skills over four distinct regions. These regions were selected to represent various atmospheric conditions and to cover large geographical domains such that sufficient observations are available for comparison. The comparison of model results with the observations revealed specific problems for each individual model. This study suggests the further improvements needed and serves as a benchmark for re-evaluations of such improvements. In general all models show deficiencies with respect to both mean concentrations and vertical gradients of important trace gases. These include ozone, CO and NO_x at the tropopause. Too strong two-way mixing across the tropopause is suggested to be the main reason for differences between simulated and observed CO and ozone values. The generally poor correlations between simulated and measured NO_x values suggest that in particular the NO_x input by lightning and the convective transport from the polluted boundary layer are still not well described by current parameterizations, which may lead to significant differences in the spatial and seasonal distribution of NO_x in the models. Simulated OH concentrations, on the other hand, were found to be in surprisingly good agreement with measured values.

Correspondence to: D. Brunner
(dominik.brunner@iac.umnw.ethz.ch)

1 Introduction

Global chemistry transport models (CTMs) and chemistry general circulation models (C-GCMs) are becoming standard tools for estimating the contribution of individual pollutant sources to trace gases on continental and global scales. One such application is the description of changes in radiative forcing due to changes in ozone caused by anthropogenic

activities. In this context the upper troposphere/lower stratosphere region (UT/LS) is particularly important because of the sensitivity of radiative forcing to the vertical distribution of ozone in particular around the tropopause (Lacis et al., 1990). A large fraction of aircraft emissions is deposited at these sensitive altitudes. As such, the impact of emissions, in particular of nitrogen oxides, by the current and future air traffic, was studied extensively in recent years and reviewed in several assessment reports (Penner et al., 1999; Brasseur et al., 1998; NASA, 1999). CTMs and C-GCMS which were playing an important role in these assessments often showed significantly differing results both in terms of background concentrations of relevant species such as NO_x and in terms of perturbations caused by aircraft emissions. These differences underline the fact that the adequate simulation of the various chemical and dynamical processes in the upper troposphere/lower stratosphere (UT/LS) region is a particularly difficult task. A high vertical model resolution is needed to adequately represent the steep concentration gradients across the tropopause and to describe mixing between the stratosphere and the troposphere. Convective processes, which are difficult to simulate, strongly affect the photochemistry of the UT region by the rapid uplift of pollutants emitted at the surface (Berntsen and Isaksen, 1999). In addition, lightning activity in deep convective clouds is an important source of nitrogen oxides to this region though its overall strength is still poorly quantified (Hauglustaine et al., 2001; Jourdain et al., 2001; Stockwell et al., 1999; Berntsen and Isaksen, 1999). Primary production of OH radicals in the UT/LS region depends not only on O_3 and H_2O levels but also on carbonyls and peroxides whose concentrations are again strongly linked to vertical transport and mixing (Jaeglé et al., 2001; Brühl et al., 2000; Prather and Jacob, 1997). Finally, due to the relatively long lifetimes and steep vertical gradients of many compounds in the UT/LS region small inaccuracies in the advection scheme may significantly affect their concentration levels (Bregman et al., 2001). The downward flux of ozone from the stratosphere to the troposphere is particularly sensitive to the formulation of transport and to upper (stratospheric) boundary conditions. Recent estimates of the global annual mean flux obtained by a number of model studies varied by at least a factor of three (Houghton et al., 2001). All these issues culminate in the substantial uncertainty of ozone budget estimates in the UT/LS region. Careful analysis of model results by comparison with observations is therefore essential.

In the framework of the EU project TRADEOFF (Aircraft emissions: Contributions of various climate compounds to changes in composition and radiative forcing – tradeoff to reduce atmospheric impact) an extensive model evaluation study was undertaken involving five state-of-the-art CTMs and two C-GCMs. Particular emphasis was given to the UT/LS region. The project TRADEOFF aims to improve our understanding of the impact of aircraft emissions on the state of the atmosphere and climate through a sequence of complementary modeling studies. Estimates of the tradeoffs of flying at different altitudes or latitudes as well as projections of future changes until the year 2050 are essential elements of the project.

The models involved in TRADEOFF have contributed to numerous previous studies on the impact of aircraft emissions (Penner et al., 1999; Brasseur et al., 1998; Grewe et al., 2001; Meijer et al., 2000; Rogers et al., 2000; Pitari and Mancini, 2001) and participated in previous evaluation studies (Bregman et al., 2001; Law et al., 2000; Rogers et al., 2000; Penner et al., 1999; Brasseur et al., 1998; Houghton et al., 2001). Recently, Law et al. (2000) compared monthly mean ozone fields of five CTMs with data from the MOZAIC program, in which ozone is measured continuously from passenger aircraft (Marenco et al., 1998). Another model intercomparison using data from MOZAIC and ozone soundings to evaluate seasonal mean profiles in the lower stratosphere was presented by Bregman et al. (2001). Grewe et al. (2001) used the extensive data set of NO_x observations from the NOXAR project, obtained onboard a Swissair B-747 aircraft over the course of more than one year, to evaluate the NO_x distributions calculated by two C-GCMs (E39/C and ULAQ GCM) and to investigate the importance of different sources. Further studies by Wang et al. (1998), Levy II et al. (1999), Emmons et al. (2000), and Bey et al. (2001) investigated the overall ability of CTMs to simulate tropospheric photochemistry, which is another focus of the present study. In these latter studies the models were compared with measurements of many different trace gases related to ozone photochemistry, obtained from a number of surface stations and aircraft measurement campaigns.

The "classical" approach for evaluating a model followed by the above studies was to aggregate the observations over specific geographical domains, altitude ranges, and time periods. Comparisons were made with corresponding model data of statistical quantities such as mean or median values and standard deviations for these aggregates. However, the model fields were usually not sampled at exactly the same times and positions as the measurements, but rather averages over entire time periods (e.g. monthly means) and domains (e.g. over a range of grid cells) were calculated since these can easily be derived from standard model output. Furthermore, model results and observations were often taken from different years assuming that interannual variations were insignificant. Here we adopt a much more direct approach by comparing each measured data point with its temporally and spatially interpolated model counterpart. However, such a "point-by-point" analysis requires simulations over the same time periods as the measurements. This is not often feasible with climate models or with CTMs that take their meteorological input from climate simulations. For two TRADEOFF models, namely ECHAM4.L39(DLR)/CHEM (subsequently named E39/C) and ULAQ, a direct comparison was therefore not applicable and hence some analysis based on the "classical" approach was included. Considering the often

Table 1. Model properties

Model	TM3	CTM2	CTM2-Gauss	TOMCAT	LMDz/INCA	ULAQ	E39/C
Operated by	KNMI	Univ. Oslo	Univ. Oslo	Cambridge	IPSL	Univ. Aquila	DLR
Model type	CTM	CTM	CTM	CTM	GCM	CTM	GCM
Meteorology	ECMWF	ECMWF	ECMWF	ECMWF	ECMWF	ULAQ-GCM	GCM
Lat×Ion resolution	3.75°×5°	T21	T21	T21	2.5°×3.75°	10°×20°	T30
Vertical levels	19 hybrid	19 hybrid	40 hybrid	31 hybrid	19	26 log-p	39 hybrid
Model top (hPa)	10 hPa	10 hPa	10 hPa	10 hPa	4 hPa	0.04 hPa	10 hPa
Transport scheme	slopes[1]	2nd order moments[2]	2nd order moments[2]	2nd order moments[2]	Van Leer[3]	Eulerian fully explicit	semi-lagr.[4]
Vertical velocities	hor. wind divergence	hor. wind divergence	hor. wind divergence	hor. wind divergence	hor. wind divergence	radiation scheme[5]	hor. wind divergence
Convection	Tiedtke[6]	Tiedtke[6]	Tiedtke[6]	Tiedtke[6]	Tiedtke[6]	Müller and Brasseur[7]	Tiedtke[6]
Lightning param.	Meijer[8]	Price[9]	Price[9]	Stockwell[10]	Jourdain[11]	Grewe[12]	Grewe[12]
Dynamical/chemical timestep (min)	120/120	60/60	60/60	30/15	15/30	60/60	30/30
Transported species	24	42	76	27	27	40[13]	13
Total species	39	52	98	49	45	70[13]	37
Gas phase + photolytic reactions	67+24	69+18	163 + 50	101+27	78+28	131 (40)	107
Heter. reactions	4 (aqueous)	2	7	0	4	10	4
Strat. chemistry	no	no	yes	no	no Cl or Br	yes	yes
NMHC chemistry	yes	yes	yes	Ethane/Propane	no	yes	no
Acetone chemistry	no	yes	yes	no	no	no	no
References	Meijer et al. (2000)	Sundet (1997); Kraabol et al. (2003)	Kraabol et al. (2003); Rummukainen et al. (1999)	Law et al. (2000)	Jourdain et al. (2001); Hauglustaine et al. (2002)	Pitari et al. (2002)	Hein et al. (2001)

[1]Russel and Lerner (1981)
[2]Prather (1986)
[3]Van Leer (1979)
[4]Williamson and Rasch (1994)
[5]Pitari (1993)
[6]Tiedke (1989)
[7]Müller and Brasseur (1995)
[8]Meijer et al. (2001)
[9]Price et al. (1996), but zonally redistributed according to convection
[10]Stockwell et al. (1999)
[11]Jourdain et al. (2001)
[12]Grewe et al. (2001)
[13]plus 44 aerosol bins

poor coverage and representativity of in-situ trace gas observations in the UT/LS region, the point-by-point strategy offers clear advantages. For example, it fully accounts for the specific meteorological situation during the measurements, and hence for the specific transport and photochemical histories of the air masses encountered. Furthermore, no averaging or resampling is required which could reduce the quality of agreement between models and observations. Finally, the method largely simplifies quantitative analysis of model performance since the two data sets are of equal size and can be easily compared using statistical methods. More details on the implementation of this method are given in Sects. 2.1 and 2.2.

A new extensive observation database was established specifically for this study, which is described in Sect. 2.2. This was necessary because previous collections as the one by Emmons et al. (2000) did not fully support our preferred point-to-point strategy which requires to use the original data files rather than gridded composites. A huge amount of model output and analysis products was generated in the course of this evaluation exercise, most of which were made accessible to the individual modeling groups through a dedicated web site (see http://www.iac.ethz.ch/tradeoff/database). Most of the observation data sets used in this study are publicly accessible through this web site. Here we can only present examples of the methods applied and highlight the main results and overall tendencies. We will focus on measurement campaigns using research aircraft which typically provide information on many different species but which have only a limited coverage in both time and space. A comparison with climatologies derived from commercial airliner measurements and ozone soundings will be presented in a complementary study by Köhler et al. (2003), which will also intercompare monthly mean trace gas distributions calculated by the models. Sect. 2.1 briefly describes the models and the experimental setup. Sections 2.2 and 2.3 provide

an overview of the structure and content of the observation database and introduce the general concepts applied for evaluating the models. Finally, an analysis of the overall model performance over four distinct regions is given in Sect. 3. In a second paper (Brunner et al., 2003) we will present a detailed comparison with two selected aircraft measurement campaigns, that is PEM-Tropics A and POLINAT/SONEX, evaluating both time-series and vertical profiles at various locations. That paper will also analyze in more detail how well the most relevant physical and chemical processes determining the distribution of different trace species are represented in the models.

2 Models, data and methods

2.1 Description of the models and simulations

Table 1 presents an overview of the key features of the models involved emphasizing their differences in terms of transport, chemistry, and model resolution. Further details can be found in the references provided in the table for each model. Five CTMs and two C-GCMs were used in TRADEOFF. The ULAQ CTM is driven by meteorological fields from the ULAQ climate-chemistry coupled model whereas the other CTMs are driven by European Center for Medium Range Weather Forecasts (ECMWF) analyses. The LMDz-INCA GCM was run in a special nudged-mode in which winds are relaxed to ECMWF analyses. Thus, except for E39/C and ULAQ the models were able to simulate the real weather conditions during individual measurement campaigns which is a prerequisite for using the point-by-point approach. The CTM2, TM3, TOMCAT and LMDz-INCA models only included tropospheric chemistry whereas both tropospheric and stratospheric chemistry were considered in CTM2-Gauss (a special version of CTM2 extending higher up into the stratosphere), E39/C, and ULAQ. Results of the SLIMCAT model, which is a stratospheric model formulated on isentropic surfaces, are not discussed here.

The following two types of model output were generated and analyzed with respect to observed trace gas distributions:

1. Point-by-point output (interpolated along-flight-path output) of the components O_3, CO, OH, HO_2, H_2O_2, H_2O, NO, NO_2, HNO_3, PAN, and Rn222. Models generating this type of output were run over the period 1995–1998 and at each simulation time step (typically of the order of 30 min) the instantaneous tracer fields were linearly interpolated to the positions of coinciding observations. The positions and times of these measurement points were provided to the modeling groups in separate tables (see Sect. 2.2 for further details).

2. Gridded monthly mean fields and standard deviations at $5^\circ \times 5^\circ$ horizontal resolution and at 30 equally (1 km) spaced vertical levels of the same trace gases as above, and in addition of net ozone production rates $P(O_3)$, lightning NO emissions, and wet HNO_3 deposition. This type of output was only used if no point-by-point data was available. Another useful application of this output is to compare it with the point-by-point data which allows to analyze the representativity of the point-by-point data and hence of the observations for a given month.

Three out of the five models with meteorological input from ECMWF, namely TM3, TOMCAT, and LMDz-INCA performed simulations for the entire time period of 1995 to 1998. The selection of a four year period was a compromise between including as many measurements as possible while keeping computation time and costs within reasonable bounds. The setup of the CTM2 model only allowed to simulate 1996 with the tropospheric version, and only 1997 with the CTM2-Gauss version of the model, respectively. E39/C performed a multi-year simulation with year 1990 climate conditions (greenhouse gas concentrations and sea surface temperatures as of year 1990) and reported monthly output of four consecutive years. ULAQ reported monthly output fields from a single year based on a year 1990 climate simulation of the driving ULAQ-GCM. Table 2 lists the emission sources used in each model. It was recommended to use a set of emission fields based on the recent IPCC Ox-Comp intercomparison exercise (Houghton et al., 2001) with some updates which have been compiled specifically for this project. These recommendations were followed with the following exceptions: In TM3 the individual NMHC contributions were somewhat different but the annual total was as recommended. In LMDz-INCA the total NO_x emissions were somewhat lower mainly due to lower fossil and bio fuel emissions (27.8 instead of 31.8 Tg(N)/yr). CO emissions due to isoprene oxidation were somewhat higher as recommended (270 instead of 165 Tg/yr). Note that no NMHCs were simulated in this version of LMDz-INCA. In the E39/C simulations generally different emissions were used, representing rather year 1990 than 2000 conditions: Total NO_x emissions were lower (38.7 instead 50.1 Tg(N)/yr) and CO concentrations were fixed at the surface according to Hein et al. (1997). As in LMDz-INCA no NMHC were simulated. Each model applied its specific implementation of a lightning NO source parameterization but the values were scaled uniformly to a global total of 5 Tg (N)/yr.

2.2 A new database for model evaluation purposes

As part of this study a new database of in-situ measurements from both aircraft and ozonesondes focusing on the UT/LS region was established. Table 3 presents an overview of the contents. Several campaigns not available in the collection of Emmons et al. (2000) were included in this new data set, that is measurements of the STREAM, JAL, CARIBIC, ACE-2, ACSOE, STRAT and POLARIS campaigns. Also,

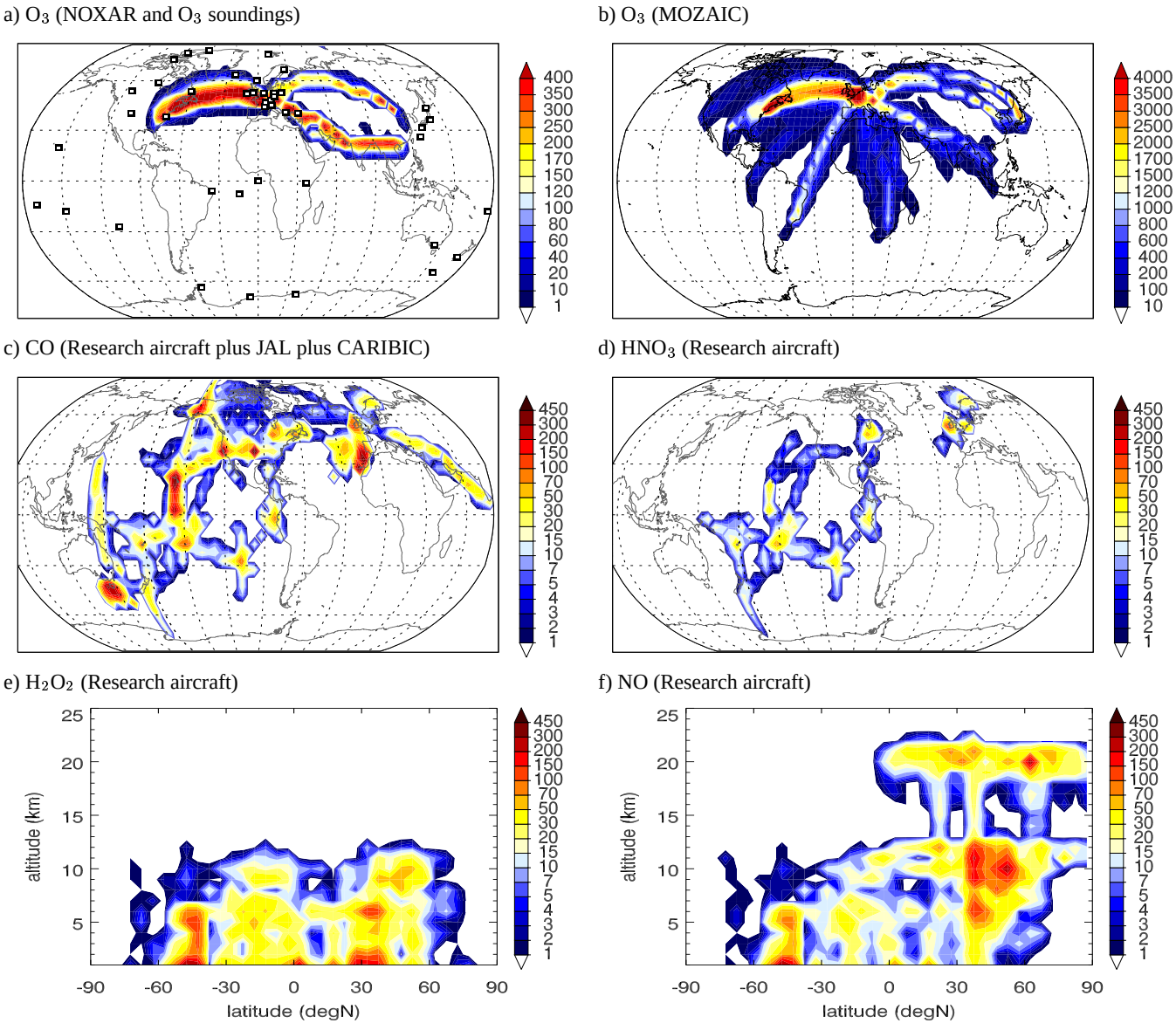

Fig. 1. Data density distributions of selected tracers in units of samples per 5° × 5° grid box (vertically integrated, panels a–d) and samples per 5° latitude × 1 km altitude grid box (zonally integrated, panels e–f), respectively. The rectangles in **(a)** are the positions of the 45 ozone sounding stations included in the database. **(c)–(f)** are distributions of selected tracers as provided by research aircraft campaigns. Also included in (c) are measurements from the two commercial aircraft measurement programs JAL (between Japan and Australia) and CARIBIC (between Germany and India).

the complete set of MOZAIC data of four consecutive years (1995–1998) was included. An important fraction is made up by the commercial aircraft measurement programs NOXAR, JAL, CARIBIC, and in particular by MOZAIC. Ozonesonde measurements obtained from the World Ozone and UV Data Center (WOUDC) and the NADIR data center at the Norwegian Institute for Air Research (NILU) contribute another important fraction. Research aircraft measurement campaigns contribute only about 2% of all data records but for many compounds this is the only reliable source of information currently available in the UT/LS region.

Figures 1a and b show the coverage of commercial aircraft measurements and ozone soundings. However, in this study we only analyze measurements from research aircraft campaigns conducted between 1995 and 1998 which are shown in Figs. 1c–f. CO was measured on nearly every research aircraft mission whereas for other species the availability is often strongly reduced. With respect to research aircraft

Table 2. Year-2000 emissions used for the TRADEOFF model runs

Species	Emission source	Strength (Tg/yr)		
		recommended[1]	LMDz-INCA	E39/C[2]
NO_x	fossil & bio fuels (∼30+1.8)	31.8 (N)	27.7	22.6
	savannah & ag-waste burning/ deforestation (3.2+1.2+2.7)	7.1	8.1	5.0
	aircraft (IPCC-TAR year 2000)	0.68	0.68	0.56
	soils	5.5	5.5	5.5
	lightning	5.0	5.0	5.0
		Total 50.1	47.0	38.7
CO	fossil fuel/domestic burning	650	650	
	deforestation/savannah & waste burning	700	514	
	vegetation(150)+oceans(50)	200	200	
		Sub-Total 1550	1364	
	CH_4 oxidation	∼800		
	isoprene oxidation[3]	∼165	270	
	terpene oxidation[3] (included in vegetation)			
	biomass burning NMHC oxidation[3]	∼30	140[4]	
	acetone oxidation[3]	∼20	20	
		Total ∼2675		
NMHC	(if considered)			
	fossil fuel/domestic burning	161 (C)		
	deforestation/savannah & waste burning	34		
	isoprene	220		
	terpene	127		
	acetone	30		
		Total 572		

[1] applies to TM3, TOMCAT, CTM2, CTM2-GAUSS and ULAQ.
[2] CO is fixed in E39/C at surface according to observations (Hein et al., 1997).
[3] use only if not treated separately as NMHC emission.
[4] includes 110 Tg(C)/yr for NMHC fossil fuel burning.

observations good coverage exists only for a small proportion of the globe whereas other important areas such as the African and Asian continents or the South Atlantic were virtually unexplored during the selected period. Nevertheless, the data set allows evaluating the models over several distinct regions under strongly differing conditions in terms of meteorology and pollutant sources (cf. Fig. 5). Figures 1e and f show that in the northern hemisphere the bulk of data was obtained in the UT/LS region between about 9 and 12 km and in midlatitudes between 30° N and 60° N. Measurements from the high-altitude aircraft ER-2 during the STRAT and POLARIS campaigns extend well into the lower stratosphere up to about 21 km. Observations in the southern hemisphere are provided mainly by the ACE-1 and PEM-Tropics A campaigns which concentrated on the Pacific region.

To exemplify Fig. 2 describes the processing of the original PEM-Tropics A campaign data files for inclusion in the database. The original files were first processed for each campaign and platform separately by averaging or rearranging all measured parameters to a common timeline and by then merging these data into a single intermediate data file per flight (labeled M1 to Mn in the figure). 6-min averaging intervals were selected for aircraft measurements and 500 m altitude intervals for the soundings. The only exceptions are PEM-Tropics A and SONEX (5-min averages), and MOZAIC (12-min averages for cruise data and 450 m altitude intervals for profiles during take-off and landing). Depending on the speed of the aircraft a 6-min time interval corresponds to a horizontal flight distance of about 50–100 km. This is still significantly smaller than the typical resolution of a global CTM which is of the order of 200–500 km. The individual flights were then combined into monthly campaign data tables which constitute the main part of the TRADEOFF database. Finally, the times and positions of observations

Table 3. Measurement campaigns and programs included in the TRADEOFF database (see http://www.iac.ethz.ch/tradeoff/database for further details)

Campaign	Periods	Trace gases	Records
MOZAIC (Marenco et al., 1998)	1995–1998	O_3, H_2O	379501
NOXAR (Brunner et al., 2001)	05 May 1995–13 May 1996, 12 Aug 1997–13 Nov 1997	NO, NO_x, O_3	44522
JAL (Matsueda et al. , 1998)	1995–1998	CO, CH_4, CO_2	1115
O_3 soundings	1995–1998	O_3, H_2O	332521
CARIBIC (Brenninkmeijer et al., 1999)	Nov 1997–Dec 1998	O_3, CO	1143
STREAM (Bregman et al., 1995)	09–14 Feb 1995 23 Nov–03 Dec 1995 22 May–01 Jun 1996 09 Mar–25 Mar 1997 27 Jun–24 Jul 1998	1995: O_3, CO, HNO_3, NO_y, acetone, N_2O 1996: O_3, H_2O, HNO_3, NO_y, acetone, N_2O 1997: O_3, CO, H_2O, HNO_3, NO_y, acet., N_2O 1998: O_3, CO, H_2O, NO, HNO_3, NO_y, acetone, H_2SO_4, CH_3CN, CO_2, N_2O (twice)	1398
STRAT	01 May–18 May 1995 19 Oct–09 Nov 1995 22 Jan–05 Feb 1996 15 Jul–10 Aug 1996 13–23 Sep 1996 02–20 Dec 1996	O_3, CO, H_2O, OH, HO_2, NO, NO_y, ethane, ethyne, N_2O, CO_2, CH_4	2331
POLINAT (Schumann et al., 2000)	21 Jun–05 Jul 1995 19 Sep–25 Oct 1997	1995: O_3, H_2O, NO, HNO_3, acetone 1996:O_3, CO, H_2O, NO, NO_x, NO_y, HNO_3, acetone, J_{NO_2}	737
ACE-1 (Bates et al., 1998)	31 Oct–22 Dec 1995	O_3, CO, OH, H_2O_2, H_2O, NO	2757
TOTE/VOTE	06 Dec 1995–19 Feb 1996	O_3, CO, H_2O, NO, NO_y, N_2O	1354
SUCCESS (Toon and Miake-Lye, 1998)	10 Apr–16 May 1996	O_3, CO, H_2O, OH, HO_2, NO, NO_y	988
PEM-Tropics A (Hoell et al., 1999)	30 Aug–05 Oct 1996	DC-8: O_3, CO, H_2O, H_2O_2, NO, NO_2, HNO_3, PAN, ethane, J_{NO_2}; P3-B: O_3, CO, H_2O, H_2O_2, OH, NO, HNO_3, ethane, J_{NO_2}	3222
ACSOE[1]	09–19 Sep 1996 02–10 Apr 1997 11 Aug– 23 Sep 1997	1996: O_3, H_2O, H_2O_2 1997: O_3, CO, H_2O_2, NO, NO_y, HCHO, J_{NO_2}	1098
POLARIS	16 Apr–15 May 1997 20 Jun–12 Jul 1997 02–25 Sep 1997	O_3, CO, H_2O, OH, HO_2, NO, NO_2, NO_y, ethane, ethyne, N_2O, CO_2, CH_4	1899
ACE-2 (Johnson et al., 2000)	16 Jun–25 Jul 1997	ARAT: O_3, H_2O; C-130: O_3, CO, H_2O_2	1847
SONEX (Thompson et al., 1999)	07 Oct–12 Nov 1997	O_3, CO, OH, HO_2, H_2O_2, H_2O, NO, NO_y, HNO_3, PAN, acetone, ethane, ethene, J_{NO_2}	1325

[1] ACSOE consisted of several sub-campaigns, including OXICOA (9–19 Sep 1996), TACIA (2–10 Apr 1997) and NARE (11 Aug–23 Sep 1997). Only data from the C-130 aircraft are included. See http://www.uea.ac.uk/\~acsoe/report.html

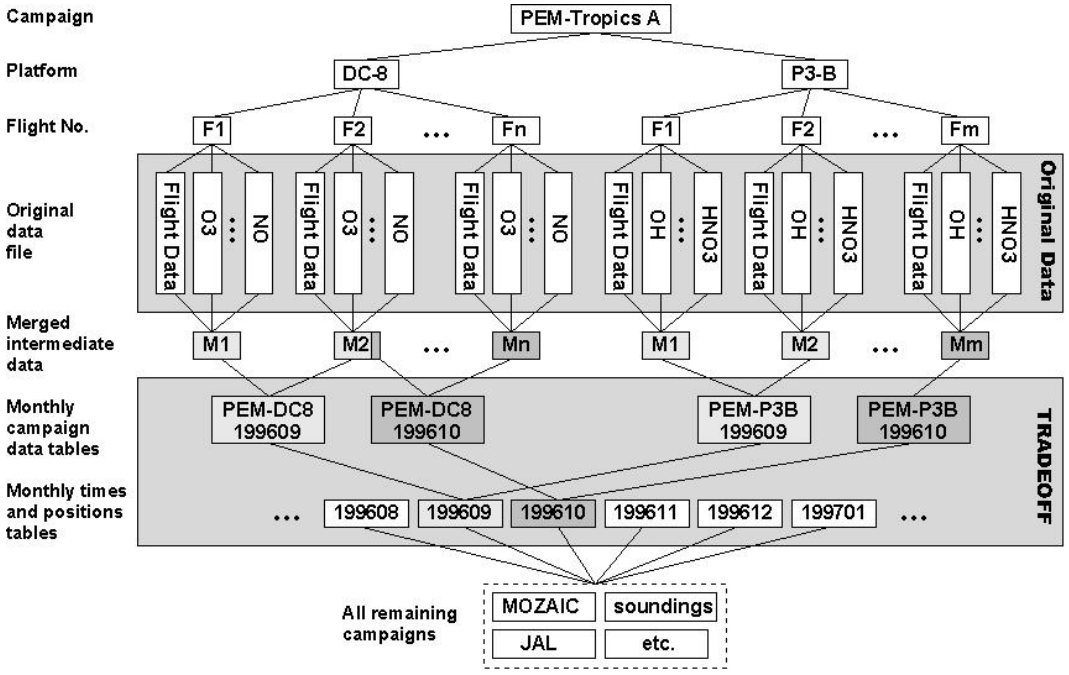

Fig. 2. Data file processing steps and structure of the TRADEOFF database. See text for further details.

from all campaigns were merged to obtain a chronologically ordered "TIMEPOS" table for each individual month, constituting the second type of tables in the database. These TIMEPOS tables were used during the model simulations to generate the output for the point-by-point comparison. Since multiple observations from different campaigns are usually available for a given time, a link between the entries in the TIMEPOS tables and the monthly campaign data tables is accomplished by an identification number. This number is chosen to be unique for a given campaign, platform, and flight, and is preserved in the point-by-point output allowing to link the model results unambiguously with the corresponding observations.

2.3 Quantitative analysis of model performance using Taylor-diagrams

Recently, Taylor (2001) presented a new type of diagram that can concisely summarize the degree of correspondence between simulated and observed fields (see Sect. 3.3 for an application of "Taylor diagrams" in this study). On this diagram the correlation coefficient R and root-mean-square (RMS) error E' between a test field f (model) and a reference field r (observations), along with the ratio of the standard deviations (σ_f and σ_r) of the two patterns are all indicated by a single point in a two-dimensional plot. Figure 3 illustrates the geometric relationship between these quantities in the diagram (see Taylor (2001) for details). The pattern

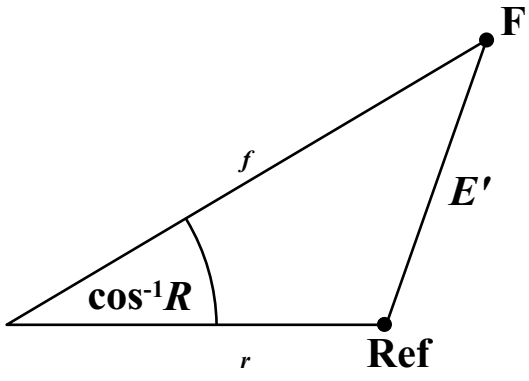

Fig. 3. Geometric relationship in a Taylor diagram between the correlation coefficient R, the root mean square (RMS) error E' and the standard deviations of the test field σ_f and reference field σ_r, respectively.

RMS E' is defined as

$$E' = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left[(f_n - \bar{f}) - (r_n - \bar{r})\right]^2} \quad (1)$$

and is thus the RMS difference between the deviations of the test and reference fields from their respective mean values $\bar{f}$ and $\bar{r}$. In the diagram the correlation coefficient is simply the cosine of the angle between the x-axis and the test point F, and the RMS difference is the linear distance between the test point and the reference point **Ref**. Hence, the point of a well

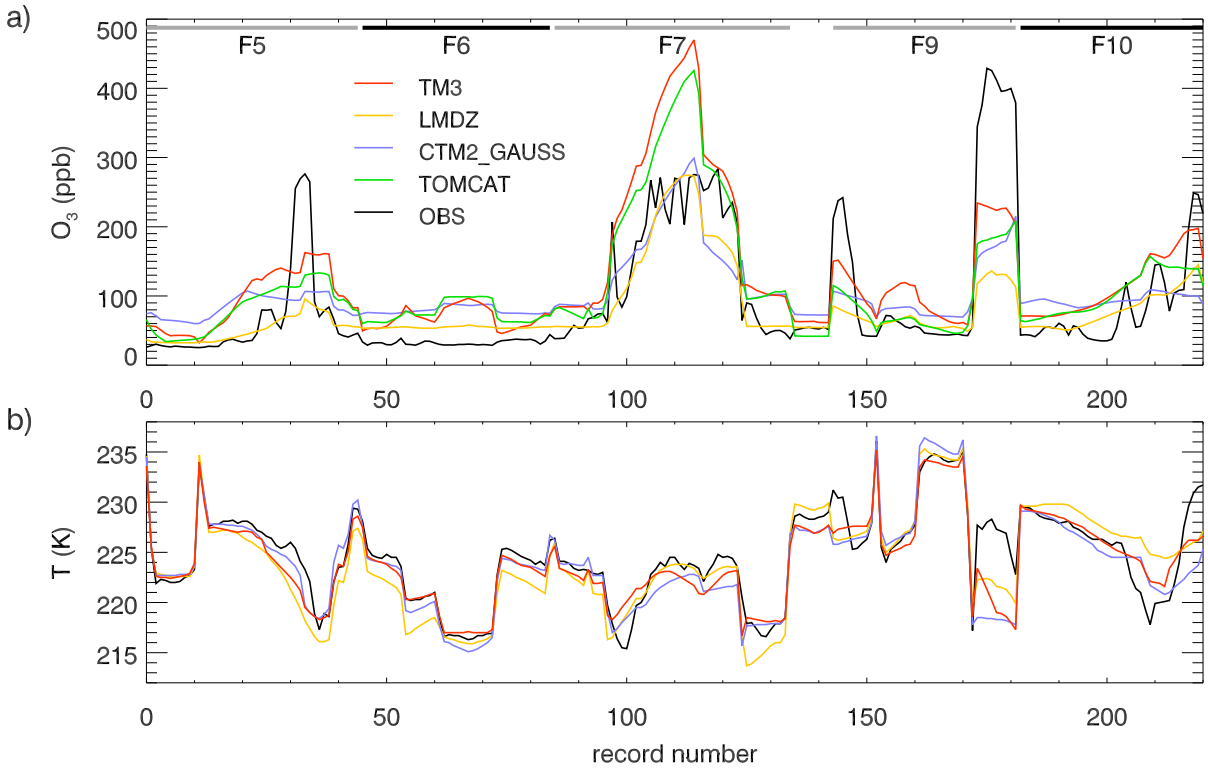

Fig. 4. Time series of **(a)** ozone and **(b)** temperature in the UT/LS region (p<300 hPa) along the flight tracks of five consecutive flights of the NASA DC-8 aircraft during the POLINAT/SONEX campaign in Oct/Nov 1997. Measured values are shown in black and interpolated model fields in color. Temperatures in the LMDz-INCA GCM have a cold bias at these altitudes. The line was therefore shifted by 5°C to match the ECMWF temperatures. Temperature is not available in the output of the TOMCAT model.

performing test field would appear near the reference point. In the following we will use standard deviations of a model field that are normalized by the observed standard deviation, denoted as $\hat{\sigma_f}$.

The correlation coefficient and the RMS difference provide complementary aspects of model performance. If, for instance, peaks of high NO concentrations caused by lightning were well represented in a model but the amplitudes of these signals were strongly underestimated, the model would exhibit a high correlation coefficient but at the same time a poor performance in terms of RMS error. On the other hand, for a given RMS value it is impossible to determine how much of the error is caused by a difference in structure and phase and how much is simply caused by errors in the amplitude of variations. Judging the overall skill of a model must therefore take into account both the correlation coefficient and the RMS error. Taylor (2001) presented several formulations for an overall skill score. Here we adopt a formulation that puts more weight on a good correlation than on a small RMS error. A good correlation suggests that fundamental processes are adequately represented in a model. We subjectively judge this here to be more important than for instance an accurate representation of the strength of emission sources which would additionally improve the skill in terms of RMS difference. We define a skill score as

$$S = \frac{4(1+R)^2}{\left(\hat{\sigma_f} + 1/\hat{\sigma_f}\right)^2 (1+R_0)^2} \quad (2)$$

where R_0 is the maximum attainable correlation. The maximum attainable correlation is limited for instance by the fact that the model fields can not fully resolve all features of the 6-min averaged observation data. Representativeness errors are a further limitation. It is not clear how representative an observation averaged along a line-shaped aircraft track is with respect to a grid volume average as provided by a model. An estimate of this error would require additional information about the real variability of the concentrations inside a model grid volume. Apart from the mean value linear gradients and second-order moments of the trace gas distribution inside a grid box are sometimes available depending on the type of advection scheme used. In principle this information could be employed to improve the interpolation from the model grid to the aircraft track. However, for simplicity we have applied the same interpolation algorithm in all models which may only account for a first order linear variation in the spatial distribution of trace gas concentrations. In order to obtain a rough estimate of the influence of the above mentioned limitations on R_0 we have investigated the correlation

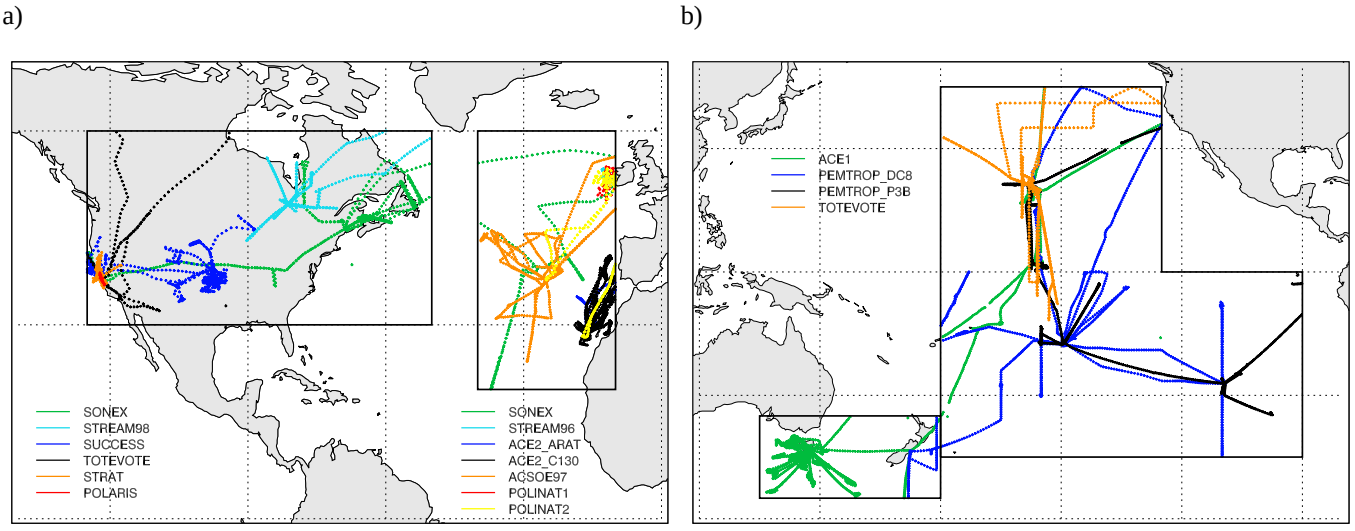

Fig. 5. The four separate domains used for the overall evaluation: North America, 30° to 60° N, 60° to 125° W; North Atlantic: 20° to 60° N, 10° to 40° W; Pacific 45° S to 45° N, 125° to 180° W (North Pacific) or 90° to 180° W (South Pacific); Tasmania: 35° to 55° S, 135° to 180° E. Overlaid are the measurement flights in different colors for the different campaigns.

between modeled and measured temperature as discussed in Sect. 3.3.

Instrument noise further reduces the maximum attainable correlation. To estimate this effect we added an artificial Gaussian noise to the point-by-point output of a particular model according to the stated instrument precision. The correlation between the original and the noisy model output then provides a measure for the influence of instrument noise on R_0.

It is important to note that the above definition of a skill score does not account for the overall bias between simulated and observed concentrations. Therefore, Taylor diagrams of model performance are presented in this study along with tables of the overall bias. In some situations the overall bias may be more indicative of a model's capabilities than a correlation coefficient. As an example, concentration time series over remote areas often resemble constants with some noisy pattern, and the correlation coefficient may be low despite the fact that the simulated concentrations lie within a few percent of the measured ones.

3 Evaluation of overall model performance in the UT/LS region

To demonstrate the capacity of the point-by-point approach Fig. 4 shows exemplarily time series of ozone and temperature along the path of 5 consecutive flights of the NASA DC-8 aircraft during the POLINAT/SONEX campaign in October/November 1997 (Singh et al., 1999; Thompson et al., 1999). Measured values are shown in black and interpolated model fields in colors. Differences between the model temperatures are very small since except for the LMDz-INCA GCM the temperature fields are taken from the driving ECMWF model. In the LMDz-INCA model only winds were relaxed to ECMWF analyses by the nudging procedure, but not temperature. LMDz-INCA has a cold bias of about 5°C at the tropopause in mid-latitudes which was corrected for in the figure. The comparison with temperature gives an impression of the limitations that different spatial resolutions between measurements and models impose on the maximum achievable agreement. The excellent agreement indicates that for the selected 5-min averages along the flight track most of the observed variability can be resolved by the models.

The variability in ozone is largely due to variations in tropopause altitudes, resulting in alternating flight stretches in the upper troposphere (O_3 below 100 ppbv) and in the lowermost stratosphere (O_3 above 100 ppbv). Variations in ozone are not as well reproduced by the models as variations in temperature, indicating that other factors apart from model resolution such as numerical diffusion must play an important role.

Occasionally, also the ECMWF analysis can severely be in error. As an example, towards the end of flight F9 the model temperatures differ from the observations by about 5 to 7°C and at the same time simulated ozone concentrations are substantially lower than observed. Hence, errors in the driving meteorological fields additionally put a limitation on the maximum achievable agreement. This limitation is likely to be most important for longlived species such as ozone and CO since their concentrations are determined by the history of air parcels over long time periods during which errors in the analyzed meteorological fields may accumulate.

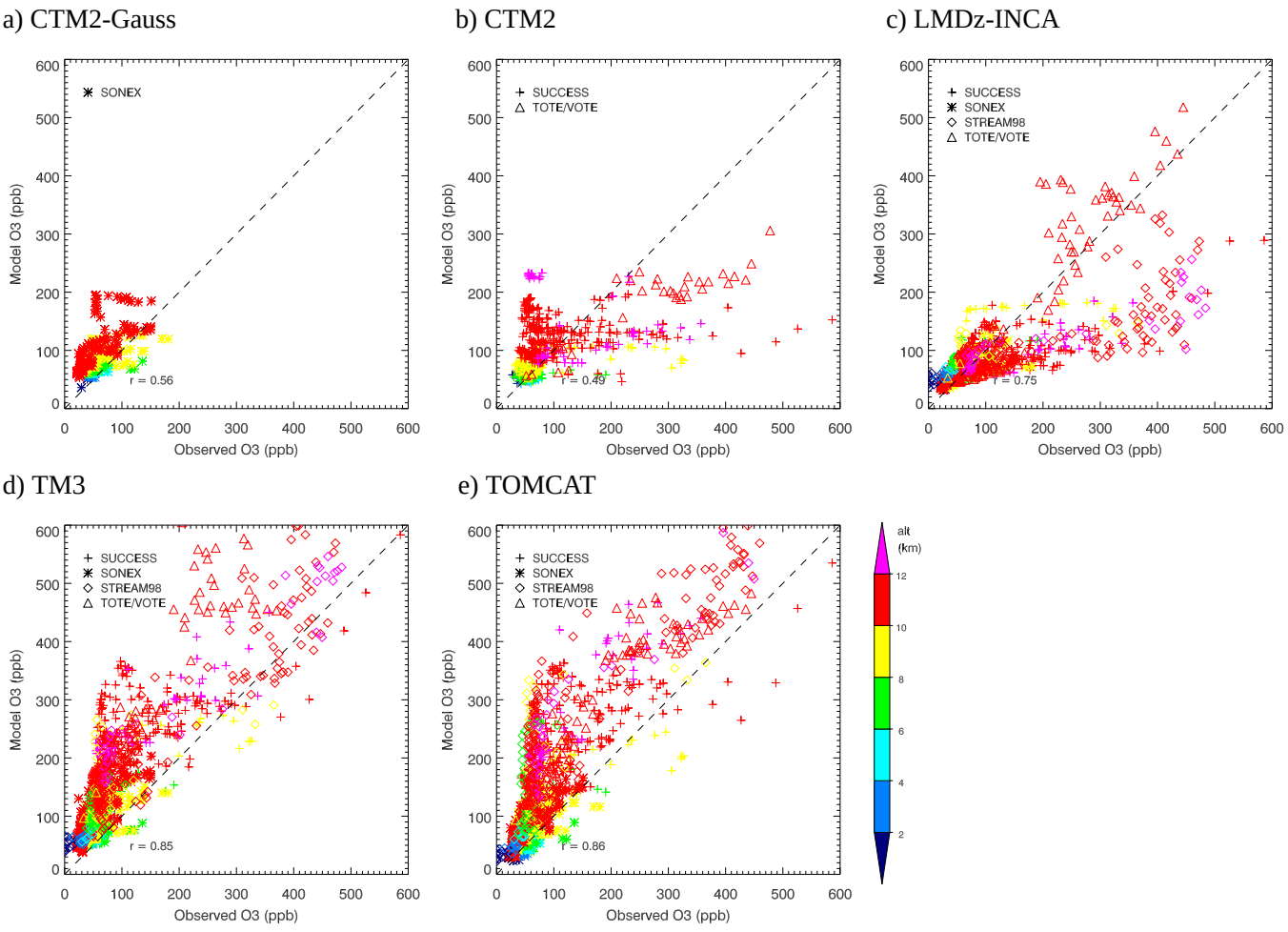

Fig. 6. Scatter plots of modelled versus measured ozone concentrations over North America. **(a)** CTM2-Gauss, **(b)** CTM2, **(c)** LMDz-INCA, **(d)** TM3, **(e)** TOMCAT. Different campaigns are represented by different symbols. Symbols are colored according to the altitude range (see the color bar). The dashed line is the 1:1 ratio.

However, we can only speculate to what extent this may influence the quality of the simulated trace gas fields since so far there have been no detailed investigations to this problem.

The kind of data shown in Fig. 4 is used in the following sections that present an evaluation of the overall model performance. The results were grouped by four different geographical regions to provide a test for the models under different meteorological conditions and remoteness of sources. The coverage of research aircraft measurements in the four regions is displayed in Fig. 5 with flights from different campaigns shown in different colors. Section 3.1 presents a few selected scatter plots of measured versus modelled trace gases for the North America region, and Sects. 3.2 and 3.3 present an analysis of average concentration biases and overall model performance in all four regions, respectively.

Finally, meridional distributions over the Pacific ocean where the measurements covered a broad range of latitudes from about 45° S to 45° N are evaluated in Sect. 3.4.

3.1 Scatter plots

Figures 6 and 7 are scatter plots between measured and simulated ozone and CO from all research aircraft measurements obtained over North America. Only the results of the LMDz-INCA, TM3, and TOMCAT models are fully comparable since output of the CTM2-Gauss and CTM2 models was only available for campaigns in 1997 (POLINAT/SONEX) and in 1996 (SUCCESS, TOTE/VOTE), respectively. The comparison with ozone shows high correlation coefficients of about 0.85 for the TM3 and TOMCAT models but they both overestimate the increase in ozone when changing from the UT into the LS by roughly a factor of two. Particularly in the range of observed concentrations between 70 to 150 ppbv the two models quite strongly overestimate ozone suggesting too strong downward mixing from the stratosphere. In contrast, lower stratospheric ozone is underestimated in the CTM2 and LMDz-INCA models. In both models the agreement is best for the TOTE/VOTE campaign which was carried out

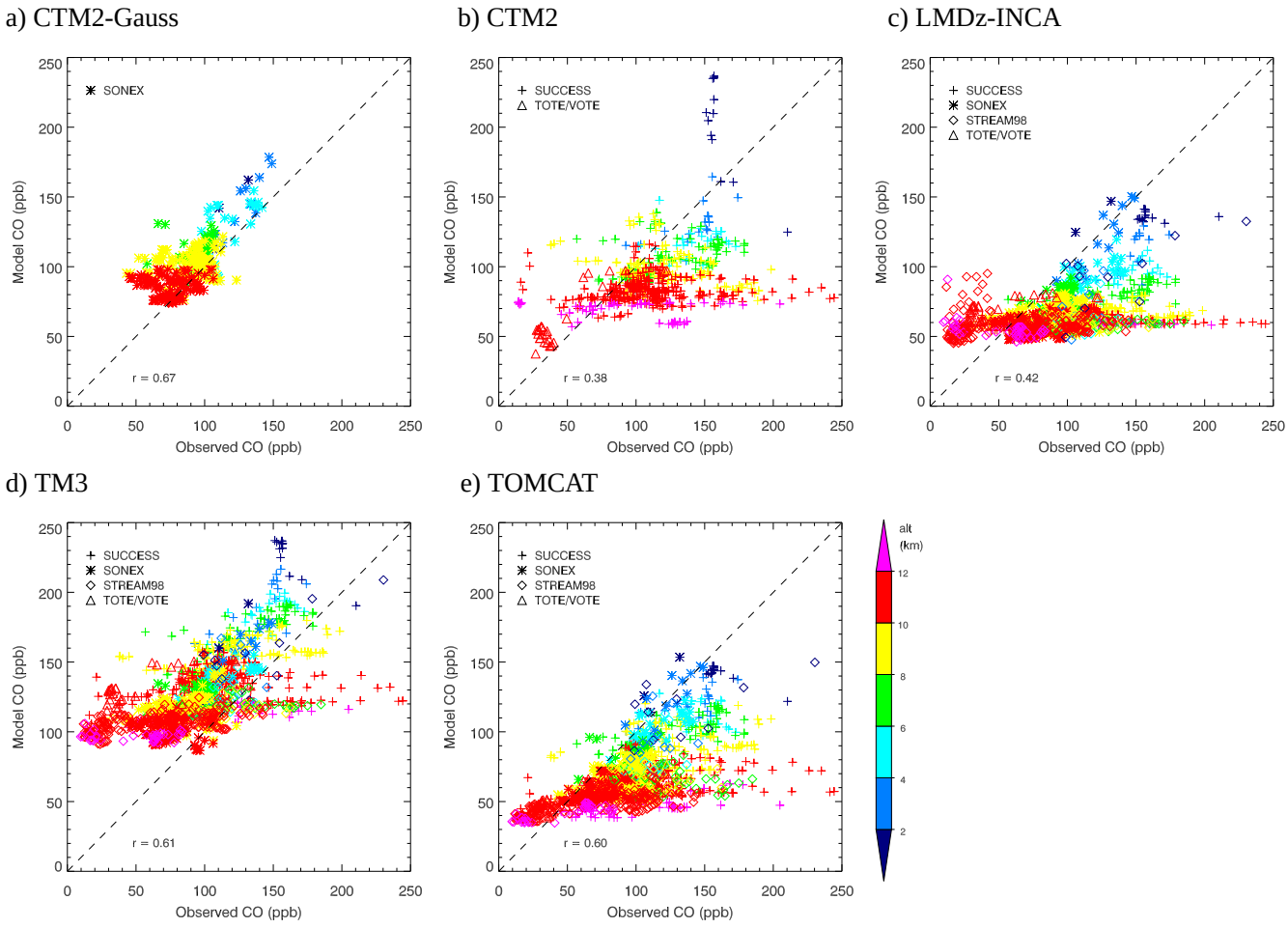

Fig. 7. Scatter plots of modelled versus measured CO concentrations over North America. See Fig. 6 for further details.

in winter. Similar to TM3 and TOMCAT the CTM2 model tends to overestimate ozone in the upper troposphere, possibly also due to too strong downward mixing. The vertical model resolution appears not to be the dominant factor in determining the ozone levels in the UT/LS region. The LMDz-INCA and TM3 models, for instance, have the same vertical resolution but they show quite a different behavior with respect to ozone. The results for TOMCAT, on the other hand, are similar to TM3 despite its substantially higher resolution (see Table 1). The CO data reveal similar problems. In all models the slope in the CO-CO correlation is too low meaning that the models tend to underestimate CO when the observed concentrations are high and vice versa. Only the CTM2-Gauss model agrees well with the observed trend but the results are limited to the POLINAT/SONEX data set which covers only a relatively small range of CO concentrations. Particularly at very low concentrations (<50 ppbv) observed in the lowermost stratosphere the models quite strongly overestimate CO. Thus, the models tend not only to mix down too much ozone from the LS into the UT, but also too much tropospheric pollutants in the opposite direction. It remains inconclusive to what extent these problems are related simply to model resolution and how much is due to numerical diffusion and other factors.

Another factor influencing the CO-CO correlation is upward transport and mixing from the boundary layer. The slope of the correlation between points in the lower troposphere (blue symbols) with typically high concentrations and points between 6 and 10 km (green, yellow) with typically lower concentrations can be used as an indicator for the intensity of vertical mixing in the models as compared to the observations. This slope looks quite reasonable in the models for the SONEX campaign but for other campaigns the scatter between simulated and observed concentrations is too large to draw a firm conclusion.

The SUCCESS measurements (represented by "+" symbols in the figures) exhibit some very high CO concentrations observed at about 10 km altitude. These high values are most likely due to rapid convective transport from the polluted boundary layer. The fact that none of the models was able to reproduce these values suggests that the convective transport during this particular event was not correctly simulated.

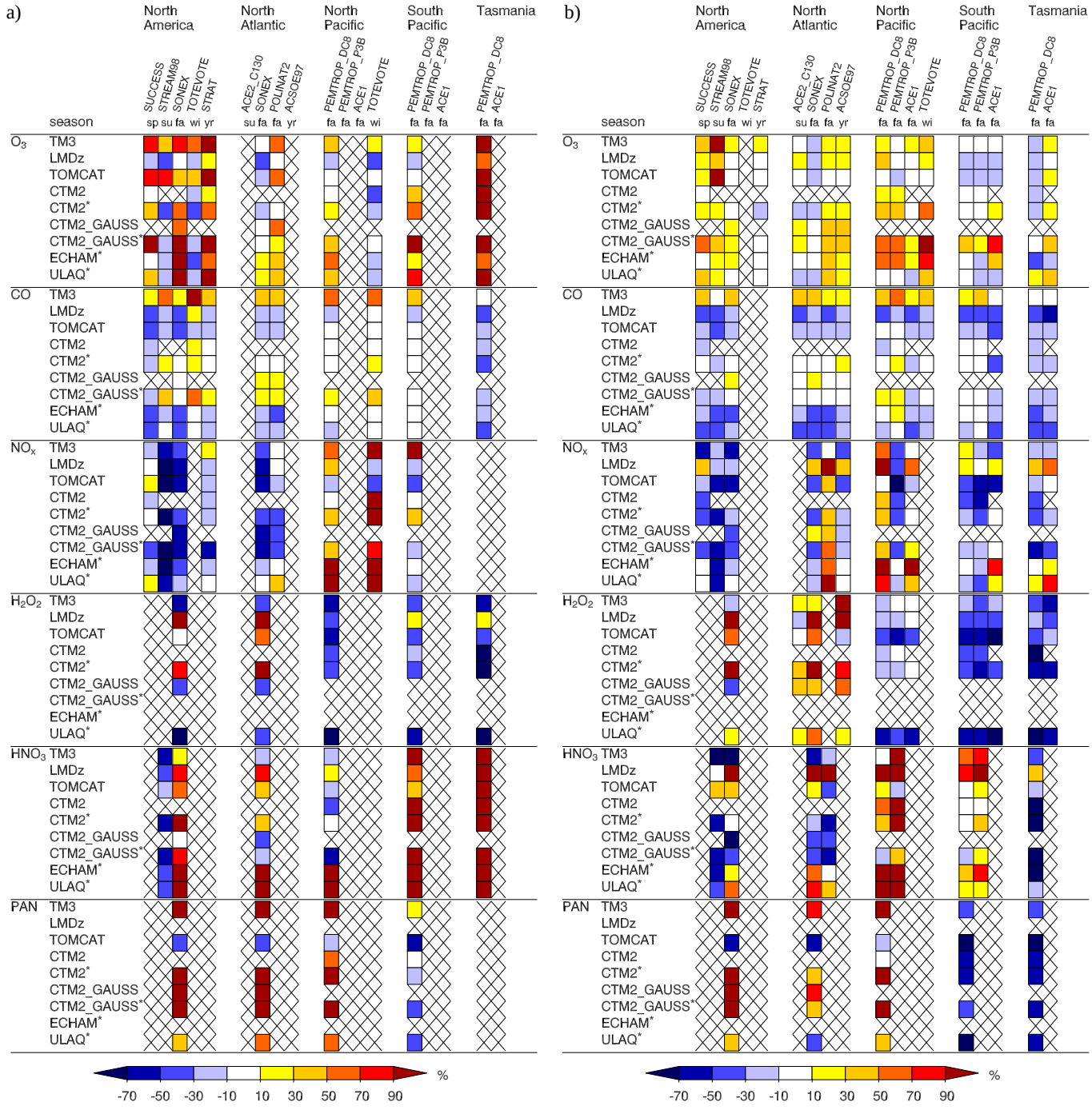

Fig. 8. Model biases $(\text{mean}_{\text{model}} - \text{mean}_{\text{obs}})/\text{mean}_{\text{obs}} * 100\%$ in the altitude range of **(a)** 300 to 170 hPa (approx. 9–13 km), and **(b)** 510 to 350 hPa (approx. 5.5 to 8 km). Biases are shown for each measurement campaign separately and grouped by the four regions. Campaign names and the corresponding season are indicated at the top (sp = spring (MAM), su = summer (JJA), fa = fall (SON), wi = winter (DJF), yr = measurements in several different seasons). The different model versions are listed in the left column. The star (*) behind a model name indicates that the values were obtained from monthly mean fields instead of point-by-point output.

3.2 Model biases

Average model biases, expressed as $(\text{mean}_{\text{model}} - \text{mean}_{\text{obs}})/\text{mean}_{\text{obs}} * 100\%$, were calculated separately for each measurement campaign and grouped by the four regions (Fig. 8). The Pacific Ocean was further separated into a North Pacific and a South Pacific domain. Figure 8a presents the results for the altitude range of 300–170 hPa (approx. 9–13 km) and Fig. 8b for lower altitudes of 510–350 hPa (approx. 5.5–8 km), respectively. The upper range covers measurements in both the UT and the LS whereas the lower range basically represents free tropospheric air with little direct influence from the stratosphere. Some measurement campaigns only covered the lower or the upper levels but not both. Missing or insufficient data are represented by crosses in the figure. Model biases are shown as colored boxes with blue colors indicating negative and yellow to red colors indicating positive model biases, respectively. Note that the above definition does not represent an average over individual biases but rather a bias of averaged model versus averaged observation data. For the observations the average was taken over all samples available for a specific campaign, region and altitude range. The corresponding model values were taken from the point-by-point output whenever possible. Since the CTM2 and CTM2-GAUSS models only covered a single year, results for these models were derived additionally from gridded monthly output fields by selecting the same months and by averaging over the same geographical domains covered by the measurement campaigns. Results for the E39/C and ULAQ models could be derived only in this way from monthly output fields.

The two figures reveal tendencies for individual models and often for entire groups of models. The TM3 model clearly simulates too high O_3 and CO concentrations at both levels and over all regions except for the lower altitude range over the South Pacific and Tasmania. The LMDz-INCA model tends to underestimate ozone in the 9–13 km altitude range but is generally in close agreement with observations at the lower levels. CO is consistently too low at both levels and over all regions. The TOMCAT model is also generally low in CO but it is closer to observed values over the Pacific ocean than at northern mid-latitudes. Remarkably, TOMCAT overestimates O_3 concentrations over North America in both altitude ranges but is close to or even below the observed values over the other regions. As seen in the previous section TOMCAT simulates too high ozone in the lowermost stratosphere. Most probably the high values over North America are due to the fact that a higher fraction of the flights took place in the lowermost stratosphere than over the other regions.

As noted before, results for the two versions of the CTM2 model were derived both from the point-by-point data and from monthly output fields. Average biases calculated in these two different ways are broadly consistent suggesting that also the comparison with monthly mean fields provides valuable information on deviations between models and observations. This is important with respect to the evaluation of the E39/C and ULAQ models for which only monthly output data is available. The CTM2-Gauss model displays a general tendency towards a positive ozone bias over all regions and at both levels whereas deviations are typically small for CTM2, in particular when only considering the more accurate point-by-point data. Considering the monthly mean data the two models behave very similar with the "Gauss version" always showing somewhat higher ozone. Also with respect to CO the two model versions behave quite similar. Deviations from observations are rather small and most of the time within 30% of the measured means.

Deviations from the observed mean O_3 and CO concentrations are similar in the E39/C and ULAQ models. However, a notable difference is evident at the lower altitudes over the North Pacific ocean where E39/C quite significantly overestimates O_3 whereas ULAQ is in close agreement with the observations. CO values tend to be too low in both models over all regions and at both levels.

All models tend to overestimate ozone and underestimate CO at the upper altitude range over Tasmania when compared to PEM-Tropics A measurements. However, it should be noted that this is only a small data sample. Too low CO concentrations are also seen in the models when compared with ACE-1 measurements, which is a much more representative data set for this region.

NO_x is the sum of NO plus NO_2. Since only few NO_2 measurements are available and the quality of the NO_2 measurements is a subject of debate (Crawford et al., 1996), we have calculated "observed NO_x" by using only measured NO concentrations and scaling these values by the NO_x to NO ratio predicted by the TM3 model. The advantage of using NO_x instead of NO alone is that NO_x concentrations are insensitive to the solar zenith angle (which only influences the partitioning between NO and NO_2) and therefore a comparison with monthly mean values is more easily possible. Observations with a NO to NO_x of less than 0.2 were omitted to exclude measurements at night and at high solar zenith angles.

The models tend to underestimate NO_x over North America at both altitude ranges and over the North Atlantic at the higher levels. Mainly elevated values are seen in the models at the lower levels over the North Atlantic when compared with POLINAT 2 and over the North Pacific when compared with PEM-Tropics A (DC-8) measurements. The TM3 model significantly overestimates NO_x at 9–13 km over the remote South Pacific. A detailed comparison with PEM-Tropics A measurements (Brunner et al., 2003) suggests a too strong lightning contribution in TM3 over this region. The LMDz-INCA model is often biased high at the lower altitudes (5.5–8 km) compared to both the measurements and the other models, in particular over the North Atlantic. This model also tends to be high in HNO_3 at those levels. Hence, excessive NO production through HNO_3 photolysis

is a likely explanation for the elevated NO_x values. In a recently revised version of the LMDz-INCA model washout of HNO_3 has been increased which would likely bring the model in better agreement with the observations. As a positive side-effect, OH concentrations, which were also found to be high in the model, were also significantly reduced (D. Hauglustaine, personal communication). Several models strongly overestimate NO over the North Pacific Ocean when compared with TOTE/VOTE measurements. Since most of the TOTE/VOTE flights were performed at night during winter, only few samples could be used to calculate NO_x concentrations and hence the representativity of this data is quite poor. However, a similar tendency is seen in the comparison with the much more representative PEM-Tropics A measurements from the DC-8 aircraft in fall. HNO_3 is much higher in all models than observed over the South Pacific. The same discrepancy was reported for other models also and several hypotheses were formulated (Wang et al., 1998; Bey et al., 2001), including insufficient wash-out of HNO_3 (Wang et al., 1998), missing heterogeneous conversion of HNO_3 to NO_x on sulfate aerosols (Chatfield, 1994) or on soot (Hauglustaine et al., 1996), overestimate of N_2O_5 hydrolysis which is suppressed if aerosols were mostly dry (Schultz et al., 2000), and removal of HNO_3 due to gravitational settling of cirrus ice crystals which is missing in the models (Lawrence and Crutzen, 1998). Our detailed comparison with PEM-Tropics A measurements (Brunner et al., 2003) also suggest too strong downward transport of HNO_3 from the lowermost stratosphere which is confirmed by similarly elevated O_3 concentrations in most models over this region.

Except for the TOMCAT model PAN is often substantially overestimated by the models. A clear exception is the PEM-Topics A measurements over the South Pacific and Tasmania. As reported by Fuelberg et al. (1999) these measurements were strongly affected by biomass burning activity. Biomass burning is an important source of PAN in the atmosphere which appears to be underestimated by the models. The closest agreement with observations is generally achieved by the ULAQ model.

Results for hydrogen peroxide reveal remarkable differences between more polluted areas such as North America and the North Atlantic and more remote regions such as the Pacific Ocean and Tasmania. H_2O_2 tends to be underestimated by all models over remote regions whereas some models overestimate its concentrations over the more polluted areas. These differences are particularly evident for the models LMDz-INCA, TOMCAT, CTM2, and ULAQ. The high H_2O_2 concentrations in the LMDz-INCA model at northern midlatitudes can be explained by the excessive water vapor simulated by this GCM. With respect to the other models the mainly too low NO concentrations over North America and the North Atlantic may contribute to the elevated H_2O_2 over these regions. The reaction of peroxy radicals with NO is in competition with the production of H_2O_2 and hence too low NO leads to excessive H_2O_2. However, differences between individual model results are inconclusive in that respect.

3.3 Taylor diagrams

The two altitude ranges, which were studied separately in the previous section were combined to a single altitude range of 510–170 hPa (approx. 5.5–13 km) in order to analyse quantitatively the model performance with respect to correlation and RMS error. The combined point-by-point model output from all campaigns available over a given region was evaluated in the form of Taylor diagrams (Fig. 9). No results can be presented here for the ULAQ and E39/C models and results for the CTM2 and CTM2-Gauss models are restricted to the years 1996 and 1997, respectively.

Isolines of skill scores S are shown in the figure as grey contours. The definition of S (see Eq. 2) includes an estimate for the maximum attainable correlation R_0. A rough estimate of this value may be obtained by considering the correlation between modeled and observed temperatures represented by the dark blue labels in the figure. Model temperatures are expected to be fairly accurate because of the assimilation of temperature observations into the driving ECMWF model. The correlation is limited by representativeness errors, the lower spatial resolution of the models as compared to the 6-min averaged observations, and errors in the meteorological analysis.

The additional influence of instrument noise on R_0 was estimated for a few individual campaigns and instruments. It was found to be most of the time significantly smaller then the influence of the errors mentioned above because instrument noise is usually very small for the 6-min averaged observations. We have therefore neglected this influence here and used the same R_0 for all trace species. This allows plotting the results for all different trace gases and different measurement campaigns in a single Taylor diagram with a single representation of skill score contours. However, this is not fully justified for instance for some of the OH and PAN measurements and for measurements of very low NO concentrations over remote regions as discussed in more detail in Brunner et al. (2003).

As expected, by far the highest scores are reached for temperature (dark blue labels) with the nudged LMDz-INCA model being almost equivalent to TM3 and CTM2-Gauss. With respect to the trace gases the highest scores are usually obtained for ozone (purple). The ozone labels for TM3 and TOMCAT often appear at normalised standard deviations greater than one meaning that these models tend to overestimate the variability. However, they usually perform best in terms of correlation resulting in a good overall skill score. CTM2 and LMDz-INCA tend to underestimate the variability in particular over regions where the stratospheric influence is strongest (as seen by high average ozone concentrations, not shown), resulting in a reduced overall skill. The same grouping of the models with respect to ozone is also apparent in Fig. 8a) showing that the TOMCAT and

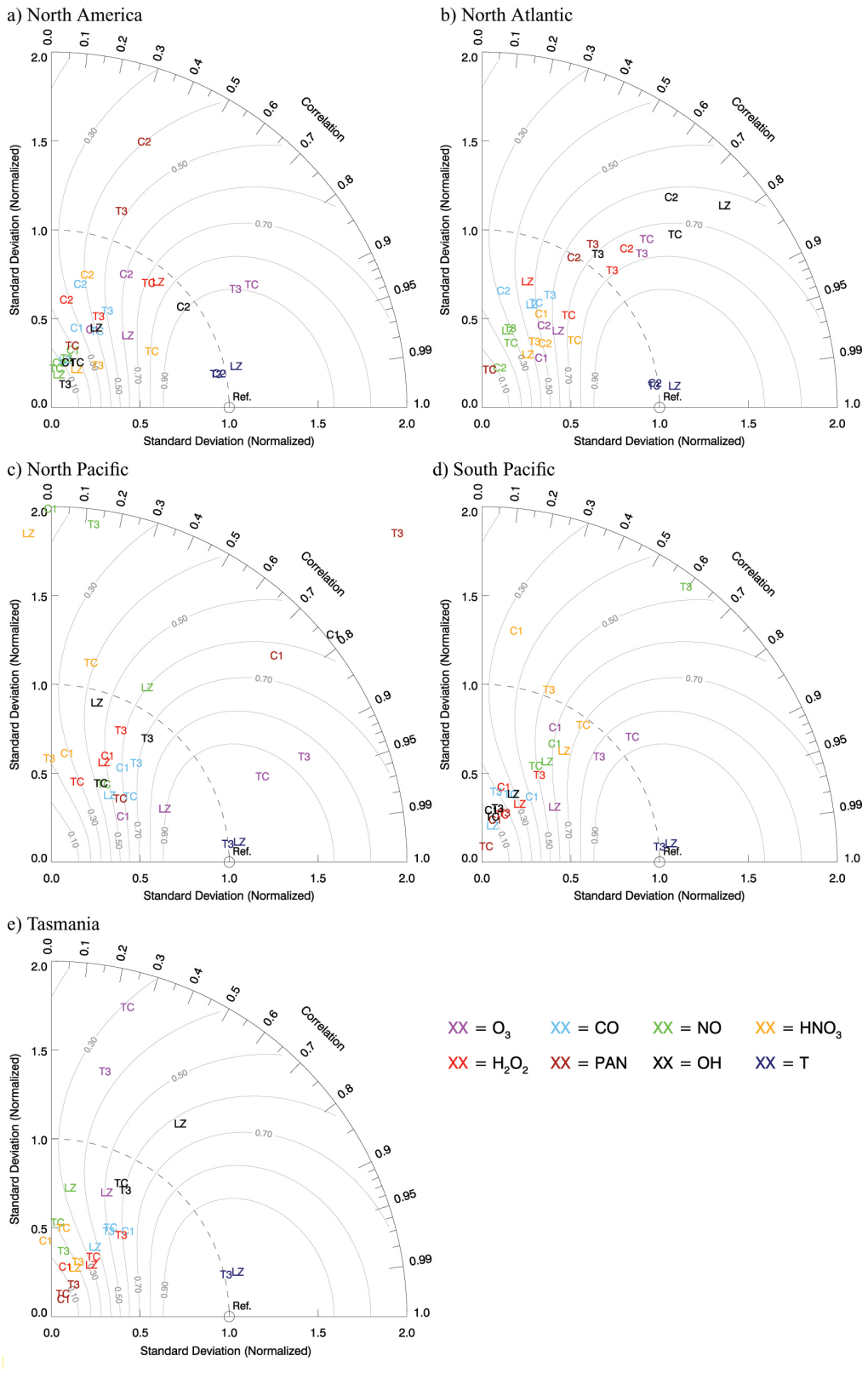

Fig. 9. Taylor diagrams for the different regions and for the altitude range of 170 hPa<p<510 hPa (approx. 5.5–13 km). Grey contour lines refer to the skill scores.

TM3 models tend to overestimate ozone in the UT/LS region whereas CTM2 and LMDz-INCA tend to slightly underestimate ozone there. All models exhibit a poor performance in terms of ozone over Tasmania.

With respect to CO (light blue) the models are grouped together closely, and, somewhat surprisingly, the skill scores are rather low over all regions. In particular over remote regions, time-series of CO often resemble constants with some noisy pattern. Therefore, the overall bias may sometimes be more indicative of the quality of the simulation than the skill score as used here. The highest scores are reached over the North Pacific and the lowest scores over the South Pacific. The unsatisfactory results for the South Pacific are most likely related to the representation of biomass burning emissions in the models. Our detailed comparison with PEM-Tropics A measurements (Brunner et al., 2003) suggests that the climatological distribution of biomass burning emissions used for these simulations is not representative for the true emissions that occurred over southern Africa, South America and Australia during fall 1996. In agreement with the results for CO the variability in PAN, another important tracer of biomass burning activity, is strongly underestimated in the models over the South Pacific.

Model performance with respect to NO is often low which points out the fact that NO is one of the most difficult species to simulate in the UT/LS region. All models tend to significantly underestimate NO variability over North America and the North Atlantic region whereas TM3 quite strongly overestimates the variability over the North and South Pacific ocean. The TM3 model also showed a significant positive bias in NO over those regions (see Fig. 8a) suggesting a too strong source of nitrogen oxides over the remote South Pacific, probably due to lightning. The TM3 model has a unique lightning parameterisation which is coupled to the intensity of convective precipitation predicted by the ECMWF model (Meijer et al., 2001), which may overestimate lightning activity over the oceans (Ernst Meijer, personal communication). Over the North Pacific also LMDz-INCA and CTM2 overstate the variability in NO along with a positive average bias. The models generally perform better with respect to HNO_3 as compared to NO over North America and the North Atlantic but they perform worse with respect to HNO_3 over the more remote North and South Pacific. This may reflect problems in simulating correctly the wash-out of HNO_3 which likely is a major factor in determining HNO_3 variability over remote regions.

In terms of H_2O_2 (red) the models perform much better over the North Atlantic than over North America. Convective activity is probably an important source of variability in H_2O_2 and OH concentrations in the upper troposphere over the continents (Prather and Jacob, 1997; Jaeglé et al., 2001). The reduced performance over North America may reflect difficulties in simulating correctly the strength, positioning and timing of convection. In agreement with this, also OH shows a strongly reduced performance over North America.

The TOMCAT model quite strongly underestimates the concentrations of PAN (not shown) and consequently its variability, resulting in a low skill score. Results of the other models for PAN are not conclusive since differences in skill scores are very large between the different regions. Obviously, the adequate simulation of PAN is a demanding task for CTMs.

3.4 Meridional distributions over the Pacific Ocean

The measurement campaigns ACE-1, PEM-Tropics A and TOTE/VOTE covered a broad range of latitudes over the Pacific Ocean from approximately 45° S to 45° N allowing to study meridional trace gas profiles over this region as shown in Fig. 10. Since we focus on the UT/LS region our analysis is restricted to the 300–170 hPa range (approx. 9–13 km). In the upper panels (Fig. 10a) the latitudinal model distributions were derived from point-by-point output. NO_x was calculated from measured NO and the ratio of NO_2 to NO predicted by the TM3 model as described in Sect. 3.2. Results of the CTM2 model were only comparable with measurements in 1996, that is with PEM-Tropics A and a fraction of the TOTE/VOTE campaign. In the lower panels (Fig. 10b) the model distributions were derived from gridded monthly output fields by selecting the same months of the year covered by the observations and by selecting the three grid levels located within the 300–170 hPa range. Both types of profiles are shown for the TM3 model in order to emphasise any differences between them.

Significant differences are seen for instance at 35–45° N where the point-by-point data show a larger contribution of stratospheric air as expressed by a strong increase in O_3 and a drop in CO concentrations. This likely explains why the climatological distributions of CTM2-Gauss, E39/C and ULAQ do not follow the observed drop in CO at these latitudes. Significant differences also exist near the lowest latitudes between 30° S and 40° S where only very few measurements are available. With respect to NO_x the two distributions differ quite strongly over the North Pacific with significantly higher NO_x values in the "climatological" monthly mean distribution comparable to the concentrations over the South Pacific. The observed distribution is much better reproduced in the point-by-point output. This underlines the importance of sampling the model data at exactly the same times and positions and hence under the same meteorological conditions, or to compare long-term means, which reflect the whole meteorological variability.

The shapes of both the O_3 and CO profiles are well reproduced by the models for which point-by-point output is available (Fig. 10a). Between 35–45° N where an important stratospheric contribution is apparent, O_3 is overestimated by TM3 and TOMCAT and underestimated by CTM2 and LMDz-INCA. This concurs with the results presented in the scatter plots of ozone over North America (Fig. 6) where TM3 and TOMCAT were found to overstate O_3 in the

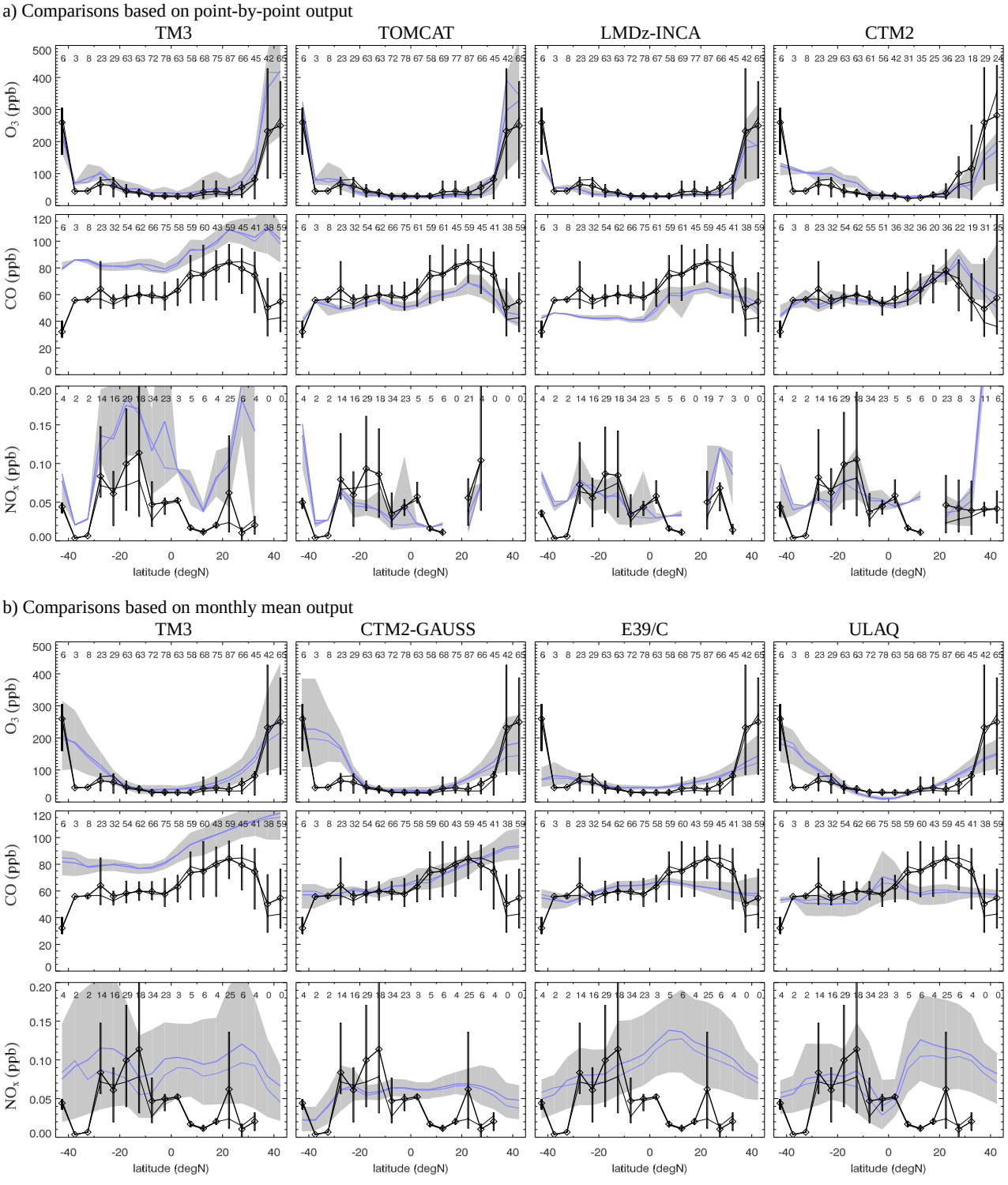

Fig. 10. Meridional distributions over the Pacific Ocean at 9–13 km altitude. **(a)** Derived from point-by-point output, **(b)** Derived from monthly mean output. The range of model values is shown as light grey shading and mean/median values as thick/thin blue lines. The range of measured values is indicated by vertical bars and mean/median values are connected by thick/thin black lines. The number of observations per 5° latitude bin is shown at the top of each panel.

lowermost stratosphere. Although the measured increase in O_3 at these latitudes is well reproduced by the models, the corresponding drop in CO is clearly underestimated, in particular by TM3 and LMDz-INCA. Since these models were run at a lower vertical resolution than TOMCAT and CTM2 this may be an issue of vertical resolution.

The measured O_3 profile shows a flat minimum around the equator which in general is well represented by the models. In the reduced data set for CTM2, both the measured and modeled minimum are shifted northward by about 10°. The E39/C model simulates too high O_3 in the tropical regions (about 45 ppbv compared to 30 ppbv) whereas the ULAQ model simulates a too strong O_3 minimum here (a minimum of about 15 ppbv).

The measurements show significantly higher CO concentrations in the Northern as compared to the Southern Hemisphere (apart from the drop at 35–45° N). This difference is also seen in the model results but the amplitude is too small in TOMCAT, E39/C, and ULAQ, suggesting a too small source or a too large sink of CO in the Northern Hemisphere. The TM3 model generally overestimates CO whereas LMDz-INCA generally underestimates the concentrations. However, the shape of the meridional profile and also the increased variability in CO in the Northern Hemisphere is well reproduced by these models. The ULAQ model shows a significant increase in CO near the equator, which is not seen in the observations. Excessively strong vertical mixing in equatorial regions is likely to produce this CO increase, because O_3 is too low at the same time. It reflects the lifting of elevated CO from continental sources and low ozone depleted in the moist tropical (marine) boundary layer (Kley et al., 1996).

The TM3/TOMCAT model overestimates/underestimates NO_x over the Pacific in both the Southern and Northern Hemisphere, in agreement with the findings presented in Fig. 8a). However, apart from this bias the main observed features are well represented. This is also true for LMDz-INCA and CTM2 which are in good agreement with the observations also in absolute terms. However, the CTM2 model simulates strongly enhanced NO_x at high latitudes not seen in the measurements. This is most likely caused by a missing sink in the model at high latitudes during winter.

With respect to NO_x the shapes of the meridional profiles of CTM2-Gauss, E39/C and ULAQ (Fig. 10b, bottom) differ quite significantly from each other, in particular in the tropics. However, the low representativeness of the measured distributions precludes the possibility of judging which profile is more realistic than the other.

4 Conclusions

We have performed a rigorous and quantitative evaluation of the performance of a number of global CTMs and C-GCMs in the UT/LS region by mainly comparing observed concentrations with model fields sampled at exactly the same times and positions as the observations. This approach is different from evaluations performed in previous studies in which typically spatially and temporally averaged fields were compared. It offers important advantages because it fully accounts for the specific meteorological conditions during the measurements. This is particularly important when comparisons are made with spatially and temporally sparse data sets such as those typically obtained from research aircraft campaigns. A new extensive observation database with better support for this type of evaluation than provided by preexisting collections has been established. We have applied a quantitative method for judging complementary aspects of model performance such as correlation coefficients and root-mean-square (RMS) errors, and we have combined this analysis with information on average biases between measurements and models. Results were analyzed separately for four different regions to test the models under varying conditions with respect to meteorology and remoteness from pollutant sources.

The main conclusions that can be drawn for individual models are:

TM3: Ozone concentrations simulated by this model are well correlated with the observations. However, TM3 significantly overestimates ozone in the lowermost stratosphere. As a consequence of this and probably also of too strong mixing across the tropopause, upper tropospheric ozone in the extratropics is also enhanced. Excessive mixing across the tropopause is also evidenced by elevated CO concentrations in the lowermost stratosphere. CO concentrations in the UT/LS region are generally biased high in the model irrespective of the geographical domain. One factor contributing to this is a deficit of OH in the UT relative to the observations. Additionally, too rapid venting of the boundary layer may transport too much CO to these altitudes. However, a sensitivity test with a less diffusive boundary layer scheme showed a reduction in upper tropospheric CO by only a few percent (Ernst Meijer, personal communication). Average NO_x concentrations are in good agreement with observations over North America and the North Atlantic, but are significantly too high over the Pacific. The unique lightning parameterization of TM3 which links lightning NO production with convective precipitation may release too much NO over the tropical ocean. Along with this nitric acid is too high in the model over the Pacific, a problem common to most CTMs in this study and also to other CTMs as reported by Wang et al. (1998) and Bey et al. (2001).

TOMCAT: Despite its higher vertical resolution TOMCAT behaves similar to TM3 with respect to ozone and CO at the tropopause, with too much O_3 and CO in the lowermost stratosphere. In contrast to TM3, the TOMCAT model is low in CO in the UT. This deficiency is more pronounced at northern mid-latitudes than in the southern hemisphere. It is unlikely that the model overestimates the sink since OH

radical concentrations are in good agreement with observations at northern midlatitudes. A more likely explanation is that TOMCAT tends to underestimate convective transport, and that the distribution of convection tends to be biased towards the tropics (M. Köhler, personal communication). NO_x tends to be low in the TOMCAT model in the UT/LS region. In the TM3 and LMDz-INCA models the vertical distribution of the lightning NO source follows the recommendations of Pickering et al. (1998) which are based on simulations with a cloud-resolving model. This is not the case in TOMCAT which likely contributes to the somewhat lower concentrations. PAN is strongly underestimated over all regions suggesting missing direct sources or missing conversion from NO_x, or both. As in all other models, hydrogen peroxide concentrations are underestimated over the North and South Pacific ocean. The reason for this discrepancy is not clear. Excessive wet removal is a possible explanation.

CTM2: In contrast to TM3 and TOMCAT, CTM2 underestimates ozone in the lowermost stratosphere. In the upper troposphere, however, ozone is often overestimated suggesting too strong downward mixing across the tropopause similar to TM3 and TOMCAT. Average CO concentrations over the different regions and also the meridional distribution over the Pacific ocean are in very good agreement with observations. The performance with respect to CO in terms of correlation coefficients and pattern RMS error, however, is similarly poor as in the other models. CO is a rather longlived species in the upper troposphere and hence time series of CO have the appearance of a constant background with a superimposed noise pattern due to recent transport events. The mean bias may therefore be more indicative for the amount of agreement with observations than the correlation coefficient. NO_x concentrations are typically 10–50% too low over North America and the North Atlantic. At high latitudes during late fall and winter, however, CTM2 strongly overestimates NO_x most probably due to missing or insufficient N_2O_5 hydrolysis on aerosols during nighttime.

CTM2-Gauss: In many respects the behavior of CTM2-Gauss (a version with stratospheric chemistry included and extending higher up in the atmosphere) resembles that of CTM2. However, since output from these two models was not available for the same year, a direct comparison is problematic. Average biases of CTM2-Gauss derived from monthly means fields versus observations show the same pattern of overestimation/underestimation as the average biases of CTM2. However, O_3 is generally somewhat higher in CTM2-Gauss in the UT/LS and NO_x somewhat lower compared to CTM2. CTM2-Gauss appears to have a similar problem of excessive NO_x at high latitudes during winter as CTM2.

LMDz-INCA: LMDz-INCA is a GCM that was run in nudged mode, in which winds were relaxed to ECMWF analyses. In this way it was possible to simulate the instantaneous weather conditions over the time period 1995 to 1998. In the extratropical UT/LS region the model is somewhat too cold and clearly too wet. Apart from this cold bias, temperature fluctuations in the model are in excellent agreement with observations (and with temperatures in the ECMWF model). The extratropical tropopause is too high in the model resulting in somewhat too low ozone concentrations at aircraft cruising altitudes. In wintertime a better agreement with observations is achieved. Scatter plots of simulated versus measured CO suggest too strong mixing across the tropopause, as seen also in the other models. LMDz-INCA shows the highest levels of OH and H_2O_2 at northern midlatitudes of all models, which is most probably related to the wet bias. Since OH concentrations determine the lifetime of many trace species including CO the concentrations of CO are consistently too low. In addition, due to missing NMHC chemistry, the secondary production of CO from NMHCs is not well represented in this model version. Nitrogen oxide concentrations are in fairly good agreement with observations at 9 to 13 km, but tend to be too high in the free troposphere at lower altitudes. In a recently revised version of the model the wet deposition of HNO_3 has been increased. It is expected that this will bring simulated HNO_3 and NO_x in better alignment with observed concentrations. As a positive side-effect OH concentrations are also reduced by approximately 15% in that model version (Didier Hauglustaine, personal communication).

ULAQ: The ULAQ model is run on a coarse grid due to the rather detailed description of microphysical processes and heterogeneous chemistry on aerosols and PSCs. The model is driven by meteorological fields from a GCM and therefore a simulation of the 1995 to 1998 period was not feasible. Hence, only a rather coarse analysis of model performance based on average biases and meridional profiles over the Pacific is possible. The ULAQ model shows no clear tendency of a positive or negative bias with respect to O_3. The agreement with observed O_3 concentrations is generally quite good and follows a pattern very similar to CTM2-Gauss. CO is generally too low in the model over all regions, but deviations from observed values are mostly within 30%. H_2O_2 is generally too low in the UT/LS region. The ULAQ model shows the best agreement in terms of average PAN concentrations of all models. The analysis of meridional trace gas distributions over the Pacific suggests that vertical mixing in the the tropics is overestimated by the model. Consequently, CO is too high and O_3, which is rapidly depleted in the tropical (marine) boundary layer, is too low in the upper tropical troposphere.

E39/C: Similar to ULAQ no point-by-point analysis was feasible since E39/C is a GCM and relaxation to meteorological analyses was not applied. Both with respect to CO and with respect to O_3 E39/C shows a similar behavior as the ULAQ model. Hence, no evidence is found for either a positive or negative bias with respect to O_3. Yet, CO tends to be generally too low in the model. NO_x concentrations are among the highest when compared to other models and are often somewhat higher than observed. However, over North

America during spring and summer, NO_x concentrations tend to be too low, as in most other models. The comparison of meridional trace gas distributions over the Pacific reveals that the difference in CO between the Southern and Northern Hemisphere is too small in E39/C. Northern hemispheric CO sources seem to be underestimated in the model. Furthermore, ozone tends to be too high in the tropical UT and too low at mid-latitudes. The underestimate of the increase in average O_3 concentrations at 9–13 km altitude when going from the subtropics to midlatitudes is probably due to a positive bias in the tropopause altitude in the extratropics.

General conclusions

Background ozone concentrations in the lowermost stratosphere differed by more than a factor of two between individual models indicating that substantial improvements of the models at these altitudes are needed. Stratospheric boundary conditions as well as vertical transport and diffusion probably have a strong impact on the model results and need to be analyzed carefully. Scatter plots of measured versus simulated ozone and CO indicate excessive two-way mixing across the tropopause. This may significantly impact the chemical environment in which aircraft emissions are released. In several models the excessive mixing appears to be responsible for elevated ozone concentrations in the upper troposphere, which in turn cause enhanced OH radical production. The abundance of HO_x radicals is a crucial factor in determining the sensitivity of ozone production to the additional release of NO_x from air traffic into the UT (Jaeglé et al., 1999). However, despite the problems in background ozone levels, OH concentrations were generally found to be in surprisingly good agreement with observed values.

Vertical model resolution could not be identified as the only major factor being responsible for excessive cross-tropopause transport. The LMDz-INCA and TM3 models, for instance, were run at the same vertical resolution but they show quite different ozone profiles across the tropopause. The results for TOMCAT, on the other hand, are similar to those for TM3, despite its substantially higher vertical resolution. With respect to the rapid drop in CO concentrations observed above the tropopause, however, models run at higher resolution perform better, although this drop was clearly underestimated even by these models. Overestimated concentrations of CO (and other trace gases of tropospheric origin) in the lowermost stratosphere poses a problem as this may result in a large sensitivity of local ozone production to aircraft NO_x emissions.

Results for NO_x differed quite significantly between the models but even more between individual campaigns. Time series that contain elevated concentrations due to fresh lightning emissions (as observed for instance on individual flights of the STREAM98 and POLINAT/SONEX campaigns, (Lange et al., 2001; Thompson et al., 1999)) cannot be reproduced by global models since their lightning parameterizations only represent grid-cell averaged sources. This is, perhaps to a lesser extent, also true for convective transport of NO_x. However, not only individual NO_x plumes, frequently observed in the upper troposphere (Brunner et al., 1998), were often missed by the models, but also campaign averaged biases revealed a larger underestimation of mean NO_x concentrations for campaigns with a substantial contribution by lightning. Even the TM3 model, which due to its specific parameterization (Meijer et al., 2001) simulates the largest lightning contribution in the upper troposphere at mid-latitudes of all models, still underestimates NO_x in these circumstances.

Acknowledgements. We would like to thank the many research and commercial aircraft measurement project leaders and managers for allowing us to use their data sets, in particular H. Schlager and U. Schumann, J.-P. Cammas and A. Marenco, C. Brenninkmeijer, H. Matsueda, S. Gaines, A. Thompson, H. Singh, B. Bregman and F. Raes. We would also like to thank Louisa Emmons for her pioneering work in this field and for the many hints on where to obtain such data sets. Finally, we highly appreciated the comments on a draft version by U. Schumann and J. Hoell, and the significant input provided by two anonymous referees. This work has been supported by the European Community grant through the project TRADEOFF (contract EVK2-CT-1999-00030).

References

Bates, T. S., Huebert, B. J., Gras, J. L., Griffiths, F. B., and Durkee, P. A.: International Global Atmospheric Chemistry (IGAC) Project's First Aerosol Characterization Experiment (ACE 1), Journal of Geophysical Research, 103, 16 297–16 318, 1998.

Berntsen, T. K. and Isaksen, I. S. A.: Effects of lightning and convection on changes in tropospheric ozone due to NO_x emissions from aircraft, Tellus (B), 51, 766–788, 1999.

Bey, I., Jacob, D. J., Yantosca, R. M., Logan, J. A., Field, B. D., Fiore, A. M., Li, Q., Liu, H. Y., Mickley, L. J., and Schultz, M. G.: Global modelling of tropospheric chemistry with assimilated meteorology: Model description and evaluation, Journal of Geophysical Research, 106, 23 073–23 096, 2001.

Brasseur, G. P., Cox, R. A., Hauglustaine, D., Isaksen, I., Lelieveld, J., Lister, D. H., Sausen, R., Schumann, U., Wahner, A., and Wiesen, P.: European scientific assessment of the effects of aircraft emissions, Atmospheric Environment, 32, 2329–2418, 1998.

Bregman, A., Van Velthoven, P. F. J., Wienhold, F. G., Fischer, H., Zenker, T., Waibel, A., Frenzel, A., Arnold, F., Harris, G. W., Bolder, M. J. A., and Lelieveld, J.: Aircraft measurements of O_3, HNO_3 and N_2O in the winter Arctic lower stratosphere during the Stratosphere-Troposphere Experiment by Aircraft Measurements (STREAM-I), Journal of Geophysical Research, 100, 11 245–11 260, 1995.

Bregman, A., Krol, M. C., Teyssèdre, H., Norton, W. A., Iwi, A., Chipperfield, M., Pitari, G., Sundet, J. K., and Lelieveld, J.: Chemistry-transport model comparison with ozone observations in the midlatitude lowermost stratosphere, Journal of Geophysical Research, 106, 17 479–17 496, 2001.

Brenninkmeijer, C. A. M., Crutzen, P. J., Fischer, H., Güsten, H., Hans, W., Heintzenberg, J., Hermann, M., Immelmann, T., Kersting, D., Maiss, M., Nolle, N., Pitscheider, A., Pohlkamp, H., Scharffe, D., and Wiedensohler, A.: CARIBIC – Civil aircraft for global measurement of trace gases and aerosols in the tropopause region, Journal of Atmospheric and Oceanic Technology, 16, 1373–1383, 1999.

Brühl, C., Pöschl, U., Crutzen, P. J., et al.: Acetone and PAN in the upper troposphere: Impact on ozone production from aircraft emissions, Atmospheric Environment, 34, 3931–3938, 2000.

Brunner, D., Staehelin, J., and Jeker, D.: Large-scale nitrogen oxide plumes in the tropopause region and implications for ozone, Science, 282, 1305–1309, 1998.

Brunner, D., Staehelin, J., Jeker, D., Wernli, H., and Schumann, U.: Nitrogen oxides and ozone in the tropopause region of the Northern Hemisphere: Measurements from commercial aircraft in 1995/1996 and 1997, Journal of Geophysical Research, 106, 27 673–27 699, 2001.

Brunner, D. Staehlin, J., Rogers, H. L., et al.: An evaluation of the performance of chemistry transport models by comparison with scientific aircraft observations, detailed comparison with the PEM-Tropics A and SONEX campaigns, in preparation, 2003.

Chatfield, R. B.: Anomalous HNO_3/NO_x ratio of remote tropospheric air: Conversion of nitric acid to formic acid and NO_x?, Geophysical Research Letters, 21, 2705–2708, 1994.

Crawford, J., Davis, D., Chen, G., et al.: Photostationary state analysis of the NO_2-NO system based on airborne observations from the western and central north pacific, Journal of Geophysical Research, 101, 2052–2072, 1996.

Emmons, L. K., Hauglustaine, D. A., Müller, J.-F., Carroll, M. A., Brasseur, G. P., Brunner, D., Staehelin, J., Thouret, V., and Marenco, A.: Data composites of airborne observations of tropospheric ozone and its precursors, Journal of Geophysical Research, 105, 20 497–20 538, 2000.

Fuelberg, H. E., Newell, R. E., Longmore, S. P., Zhu, Y., Westberg, D. J., Browell, E. V., Blake, D. R., Gregory, G. L., and Sachse, G. W.: A meteorological overview of the Pacific Exploratory Mission (PEM) Tropics period, Journal of Geophysical Research, 104, 5585–5622, 1999.

Grewe, V., Brunner, D., Dameris, M., Grenfell, J. L., Hein, R., Shindell, D., and Staehelin, J.: Origin and variability of upper tropospheric nitrogen oxides and ozone at northern mid-latitudes, Atmospheric Environment, 35, 3421–3433, 2001.

Grewe, V., Dameris, M., Fichter, C., and Sausen, R.: Impact of aircraft NO_x emissions, Meteorologische Zeitschrift, 3, 177–186, 2002.

Hauglustaine, D. A., Ridley, B. A., Solomon, S., Hess, P. G., and Madronich, S.: HNO_3/NO_x ratio in the remote troposphere during MLOPEX2: Evidence for nitric acid reduction on carbonaceous aerosols, Geophysical Research Letters, 23, 2609–2612, 1996.

Hauglustaine, D. A., Emmons, L., Newchurch, M., Brasseur, G., Takao, T., Matsubara, K., Johnson, J., Ridley, B., Stith, J., and Dye, J.: On the role of lightning NO_x in the formation of tropospheric ozone plumes: A global model perspective, Journal of Atmospheric Chemistry, 38, 277–294, 2001.

Hein, R., Crutzen, P. J., and Heimann, M.: An inverse modeling approach to investigate the global atmospheric methane cycle, Global Biogeochemical Cycles, 11, 43–76, 1997.

Hein, R., Dameris, M., Schnadt, C., Land, C., Grewe, V., Köhler, I., Ponater, M., Sausen, R., Steil, B., Landgraf, J., and Brühl, C.: Results of an interactively coupled atmospheric chemistry-general circulation model: Comparison with observations, Annales Geophysicae, 19, 435–457, 2001.

Hoell, J. M., Davis, D. D., Jacob, D. J., Rodgers, M. O., Newell, R. E., Fuelberg, H. E., McNeal, R. J., Raper, J. L., and Bendura, R. J.: Pacific Exploratory Mission in the tropical Pacific: PEM-Tropics A, August–September 1996, Journal of Geophysical Research, 104, 5567–5583, 1999.

Houghton, J. T., Ding, Y., Griggs, D. J., Noguer, M., van der Linden, P. J., and Xiaosu, D. (Eds): IPCC Third Assessment Report: Climate Change 2001: The Scientific Basis, Cambridge University Press, UK, 2001.

Jaeglé, L., Jacob, D. J., Brune, W. H., Faloona, I. C., Tan, D., Kondo, Y., Sachse, G. W., Anderson, B., Gregory, G. L., Vay, S., Singh, H. B., Blake, D. R., and Shetter, R.: Ozone production in the upper troposphere and the influence of aircraft during SONEX: Approach to NO_x saturated conditions, Geophysical Research Letters, 26, 3081–3084, 1999.

Jaeglé, L., Jacob, D. J., Brune, W. H., and Wennberg, P. O.: Chemistry of HO_x radicals in the upper troposphere, Atmospheric Environment, 35, 469–489, 2001.

Johnson, D. W., Osborne, S. R., Wood, R., Suhre, K., Andreae, M. O., Johnson, R., Businger, S., Quinn, P. K., Bates, T., Durkee, P., Johnson, H., Russell, L. M., Noone, K., Glantz, P., Bandy, B., O'Dowd, C., Rapsomanikis, S., and Rudolph, J.: An overview of the Lagrangian experiments undertaken during the North Atlantic Regional Aerosol Characterisation Experiment (ACE - 2), Tellus (B), 52, 290–320, 2000.

Jourdain, L., Hauglustaine, D. F., and Hourdin, F.: The global distribution of lightning NO_x simulated on-line in a general circulation model, J. Phys. Chem. Earth (C), 26, 585–591, 2001.

Kley, D., Crutzen, P. J., Smit, H. G. J., Vömel, H., Oltmans, S. J., Grassl, H., and Ramanathan, V.: Observations of near-zero ozone concentrations over the convective pacific: Effects on air chemistry, Science, 274, 230–233, 1996.

Köhler, M. O., Rogers, H. L., Pyle, J. A., et al.: Model intercomparison and validation with observations made aboard commercial aircraft, Atmospheric Environment, in preparation, 2003.

Lacis, A. A., Wuebbles, D. J., and Logan, J. A.: Radiative forcing of climate by changes in the vertical distribution of ozone, Journal of Geophysical Research, 95, 9971–9981, 1990.

Lange, L., Hoor, P., Helas, G., Fischer, H., Brunner, D., Scheeren, B., Williams, J., Wong, S., Wohlfrom, K.-H., Arnold, F., Ström, J., Krejci, R., Lelieveld, J., and Andreae, M. O.: Detection of lightning-produced NO in the midlatitude upper troposphere during STREAM 98, Journal of Geophysical Research, 106, 27 777–27 785, 2001.

Law, K. S., Plantevin, P. H., Thouret, V., et al.: Comparison between global chemistry transport model results and measurements of ozone and water vapor by airbus in-service aircraft (MOZAIC) data, Journal of Geophysical Research, 105, 1503–1525, 2000.

Lawrence, M. G. and Crutzen, P. J.: The impact of cloud particle gravitational settling on soluble trace gas distributions, Tellus (B), 50, 263–289, 1998.

Levy II, H., Moxim, W. J., Klonecki, A. A., and Kasibhatla, P. S.: Simulated tropospheric NO_x: Its evaluation, global distribution and individual source contributions, Journal of Geophysical Re-

search, 104, 26 269–26 306, 1999.

Marenco, A. Thouret, V., Nedelec, P., et al.: Measurement of ozone and water vapor by airbus in-service aircraft: The MOZAIC airborne program, an overview, Journal of Geophysical Research, 103, 25 631–25 642, 1998.

Matsueda, H., and Inoue, H. Y.: Measurements of atmospheric CO_2 and CH_4 using a commercial airliner from 1993 to 1994, Atmospheric Environment, 30, 1647–1655, 1996.

Matsueda, H., Inoue, H. Y., Sawa, Y., Tsutsumi, Y., and Ishii, M.: Carbon monoxide in the upper troposphere over the western Pacific between 1993 and 1996, Journal of Geophysical Research, 103, 19 093–19 110, 1998.

Meijer, E. W., van Velthoven, P. F. J., Thompson, A. M., Pfister, L., Schlager, H., Schulte, P., and Kelder, H.: Model calculations of the impact of NO_x from air traffic, lightning, and surface emissions, compared with measurements, Journal of Geophysical Research, 105, 3833–3850, 2000.

Meijer, E. W., van Velthoven, P. F. J., Brunner, D., and Kelder, H.: Improvement and evaluation of the parameterisation of nitrogen oxide production by lightning, J. Phys. Chem. Earth (C), 26, 577–583, 2001.

Müller, J.-F. and Brasseur, G., A three-dimensional transport model of the global troposphere, Journal of Geophysical Research, 100, 16 445–16 490, 1995.

NASA: Atmospheric Effects of Aviation, A Review of NASA's Subsonic Assessment Project, National Academy Press, Washington, D.C., 1999.

Penner, J. E., Lister, D., Griggs, D., Docken, D., and MacFarland, M., (Eds): IPCC Special Report on Aviation and the Global Atmosphere, Cambridge University Press, New York, 1999.

Pickering, K., Wang, Y., Wei-Kuo, T., Price, C., and Müeller, J.-F.: Vertical distributions of lightning NO_x for use in regional and global chemical transport models, Journal of Geophysical Research, 103, 31 203–31 216, 1998.

Pitari, G.: A numerical study of the possible perturbation of stratospheric dynamics due to Pinatubo aerosols: implications for tracer transport., Journal of Atmospheric Science, 50, 2443–2461, 1993.

Pitari, G. and Mancini, E.: Climatic impact of future supersonic aircraft: Role of water vapour and ozone feedback on circulation, J. Phys. Chem. Earth (C), 26, 571–576, 2001.

Prather, M. J.: Numerical advection by conservation of second-order moments, Journal of Geophysical Research, 91, 6671–6681, 1986.

Prather, M. J. and Jacob, D. J.: A persistent imbalance in HO_x and NO_x photochemistry of the upper troposphere driven by deep tropical convection, Geophysical Research Letters, 24, 3189–3192, 1997.

Price, C., Penner, J., and Prather, M., NO_x from lightning, part I: Global distribution based on lightning physics, Journal of Geophysical Research, 102, 5929–5941, 1997.

Rogers, H. L., Chipperfield, M. P., Bekki, S., and Pyle, J. A.: The effects of future supersonic aircraft on stratospheric chemistry modeled with varying meteorology, Journal of Geophysical Research, 105, 29 359–29 369, 2000.

Russel, G. L. and Lerner, J. A.: A new finite-differencing scheme for the tracer transport equation, Journal of Applied Meteorology, 20, 1483–1498, 1981.

Schultz, M. G., Jacob, D. J., Bradshaw, J. D., Sandholm, S. T., Dibb, J. E., Talbot, R. W., and Singh, H. B.: Chemical NO_x budget in the upper troposphere over the tropical south pacific, Journal of Geophysical Research, 105, 6669–6679, 2000.

Schumann, U., Schlager, H., Arnold, F., Ovarlez, J., Kelder, H., Hov, Ø., Hayman, G., Isaksen, I. S. A., Staehelin, J., and Whitefield, P. D.: Pollution from aircraft emissions in the North Atlantic flight corridor: Overview on the POLINAT projects, Journal of Geophysical Research, 105, 3605–3631, 2000.

Singh, H. B., Thompson, A. M., and Schlager, H.: The 1997 SONEX aircraft campaign and coordinated POLINAT 2 activity: Overview and accomplishments, Geophysical Research Letters, 26, 3053–3056, 1999.

Stockwell, D., Giannakopoulos, C., Plantevin, P.-H., Carver, G. D., Chipperfield, M. P., Law, K. S., Pyle, J. A., Shallcross, D. E., and Wang, K.-Y.: Modelling NO_x from lightning and its impact on global chemical fields, Atmospheric Environment, 33, 4477–4493, 1999.

Taylor, K. E.: Summarizing multiple aspects of model performance in a single diagram, Journal of Geophysical Research, 106, 7183–7192, 2001.

Thompson, A. M., Sparling, L. C., Kondo, Y., Anderson, B. E., Gregory, G. L., and Sachse, G. W.: Perspectives on NO, NO_y, and fine aerosol sources and variability during SONEX, Geophysical Research Letters, 26, 3073–3076, 1999.

Tiedtke, M.: A comprehensive mass flux scheme for cumulus parameterization in large-scale models, Monthly Weather Review, 117, 1779–1800, 1989.

Toon, O. B. and Miake-Lye, R. C.: Subsonic Aircraft: Contrail and Cloud Effects Special Study (SUCCESS), Geophysical Research Letters, 25, 1109–1112, 1998.

van Leer, B.: Towards the ultimate conservative difference scheme, V: A second order sequel to Godunov's method, Journal of Computational Physics, 32, 101–136, 1979.

Wang, Y., Logan, J. A., and Jacob, D. J.: Global simulation of tropospheric O_3-NO_x-hydrocarbon chemistry – 2. model evaluation and global ozone budget, Journal of Geophysical Research, 103, 10 727–10 755, 1998.

Williamson, D. L., and Rasch, P. J.: Water vapor transport in the NCAR CCM2, Tellus (A), 46, 34–51, 1994.